



Proceedings of the 5th International
Information Technology Conference
IITC 2003

Colombo
Sri Lanka

ENTER →



Towards an ICT enabled Society

Proceedings of the 5th International Information
Technology Conference
IITC 2003

Colombo
Sri Lanka

V.K. Samaranayake, A.R. Weerasinghe and P. Wimalaratne (Eds)

1st – 7th December 2003
BMICH,
Colombo.

Organized by the Infotel Lanka Society Ltd.



Disclaimer

The views expressed in the papers published in these proceedings are solely those of the authors and they do not necessarily represent the views of the Infotel Lanka Society Ltd.

Proceedings of the 5th International Information Technology Conference 2003

Conference Website: <http://www.iitc.lk>

This work is subject to copyright. All rights are reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Code of Intellectual Property Act of Sri Lanka 2003.

ISBN: 955-8974-00-5

Copyright © 2003 Infotel Lanka Society Ltd.

Printed from camera ready copy supplied by the University of Colombo School of Computing

Preface

The International Information Technology Conferences (IITC) began in conjunction with the government of Sri Lanka declaring the year 1998 as the Year of Information Technology. In the ensuing years IITC has made enormous efforts to maintain an unbroken run despite the many changes affecting Sri Lanka. This determination has seen the Conference evolve into a key event in the ICT calendar of the country. This can be amply evidenced by the increasing levels of sponsorship over the years by the industry.

The IITC focuses on ICT research directions and thus complements other local conferences such as the National IT Conference organized annually by the CSSL which have an industry applications orientation. As such the IITC attempts to showcase more of the future trends in ICT by attracting papers on current cutting-edge research.

This year's conference themed *Towards an ICT enabled Society* features 26 internationally refereed papers on various areas of ICT research ranging from the emerging Bio Informatics field, to digital government, e-Commerce and m-Commerce, Computer and Network Security, Human Language Technology and Localization, User Interface and Web Technologies, Visual Computing and High Performance & Concurrent Computing.

The IITC workshops and tutorials program has also become popular over the years owing to the high quality and relevancy of the topics covered in it to the local industry. This is a forum for technology experts to provide the industry with updates on technologies that are emerging in the global scene. This year's workshop/tutorials cover areas such as Software Localization through UNICODE, Bio Informatics, Computational Intelligence, e-Learning Design, Computing for the Blind, Developing Secure Applications, Broadband Access Options, Web Services, Business Process Re-engineering and Implementing Wireless Networks.

The Conference and its associated workshop/tutorial program provide a unique opportunity for initiating and enhancing industry-research partnerships and collaboration both locally and with potential international partners. This is very much at the center of the overall long-term objectives of the IITC and is increasingly becoming crucial in the development of Sri Lanka as an international player of repute as envisaged through the government's e-Sri Lanka Program.

We would like to acknowledge the ready support of the local ICT industry in coming forward with generous sponsorship of the entire IITC 2003 program and are convinced that theirs is a significant investment towards the promotion and realization of the e-Sri Lanka initiative.

We also thank all paper authors and presenters as well as workshop and tutorial resource persons for helping make IITC 2003 a fruitful experience for all participants.

Finally we wish to record our genuine appreciation of all conference and workshop/ tutorial delegates without whose presence and participation IITC 2003 would not be able to achieve its objectives.

Main Organizing Committee



Table of Contents

Evaluation of Session Identification Heuristics in Web Usage Mining	1
<i>Amithalal Caldera and Yogesh Deshpande</i>	
KANSEI Information Processing for Perception of Vowel Sounds	8
<i>Chandrajith Ashuboda Marasinghe, Ajith P. Madurapperuama, Stephen G. Lambacher, William L. Martens, Michael Cohen, and Susantha Herath</i>	
Chip-based National ID Card - Characteristics & Feasibility	14
<i>Chih-Chun Chang</i>	
Mobile Technologies, providing new possibilities in Customer Relationship Management.....	23
<i>D.Arunatileka</i>	
Optimized Web Information Retrieval with Fuzzy Logic	32
<i>Harshana Liyanage, G.E.M.D.C. Bandara</i>	
Document Management Techniques & Technologies	40
<i>Joseph P. Sathiadas, G.N. Wikramanayake</i>	
Public Key Infrastructure Security and Interoperability Testing and Evaluation.....	49
<i>Job Asheri Chaula, Louise Yngström, Stewart Kowalski</i>	
Authorization System in Open Networks Based on Attribute Certificates.....	59
<i>Jeffy Mwakalinga¹, Eric Rissanen, Sead Muftic</i>	
Web Based Generic File Protection System	68
<i>Bimali Arsakularatne, Kasun De Zoysa, Rasika Dayarathna</i>	
Framework for modelling of tacit knowledge-Case study using Ayurvedic domain	76
<i>D.S. Kalana Mendis, Asoka S. Karunananda, U. Samarathunga</i>	
Image Coding Using the Self-Similarity of Wavelet High-Frequency Components	83
<i>S.Selvarajan, N.D.Kodikara</i>	
Comparative Study on Security Issues of Wireless LAN	91
<i>Md. Enamul Haque, Meeraz Ahmed Saadi</i>	
Some aspects of coordination paradigm in concurrent programming	95
<i>D N Ranasinghe</i>	

An Approach to eTransform Enterprises in Developing Countries (A Case Study of an Enterprise in the Ceramics Sector in Sri Lanka).....	101
<i>Mahesha Kapurubandara, Shiromi Arunatileka, Prof. Athula Ginige</i>	
Web Site Visualisation as a User Navigation Aid	111
<i>Shantha Jayalal Pearl Brereton Chris Hawksley</i>	
Data Protection Law an E-Business and E-Government Perception	122
<i>Prathiba Mahanamahewa</i>	
Hybrid Ant Colonies for Parallel Algorithms	129
<i>W R M U K Wickramasinghe, D N Ranasinghe</i>	
A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation...	136
<i>Ruvan Weerasinghe</i>	
A Tool for the Management of ebXML Resources	142
<i>S.S. Sooriarachchi, G.N. Wikramanayake, G.K.A. Dias</i>	
Asymmetry in Facial Expressions: 3D Analysis using Laser Rangefinder Systems.....	152
<i>Pujitha Gunaratne, Nihal Kodikara, Yukio Sato</i>	
eMoney Order System: The Smart way to Pay Online.....	161
<i>Kasun De Zoysa, Rasika Dayarathna</i>	
A Notarization Authority for the Next Generation of E-Mail Systems	166
<i>Hiran Ekanayake, Kasun De Zoysa, Rasika Dayarathna</i>	
Learning Patterns: Towards the Personalization of E-Learning	171
<i>Dr. K. P. Hewagamage and R. S. Lekamarachchi</i>	
The Effectiveness of Digital Government as a Tool for Improved Service Delivery	180
<i>Mehdi Asgarkhani</i>	
Non-intentional Cooperative Behaviour for an Agent Based Intelligent Environment..	188
<i>R.A. Chaminda Ranasinghe, Ajith P. Madurapperuma</i>	
Engineering Optimisation with Evolutionary Computation	194
<i>Asanga Ratnaweera , Saman K. Halgamuge, Harry C. Watson</i>	
Workshops/Tutorials.....	204



Evaluation of Session Identification Heuristics in Web Usage Mining

Amithalal Caldera and Yogesh Deshpande
School of Computing and Information Technology
College of Science, Technology and Engineering
University of Western Sydney
PO Box 1797, Penrith South DC, NSW 1797, Australia

h.caldera@uws.edu.au, y.deshpande@uws.edu.au

Abstract

Web Usage Mining (WUM) is the discovery of interesting knowledge from Web server logs. The access log files of a Web server contain a lot of details about users' on-site behaviour. The validity of WUM depends on the accurate identification of user sessions implicitly recorded in these logs. In some applications, a user may be explicitly identified through user authentication. However, in general, the Web logs do not contain a user id and separate user sessions have to be inferred through heuristics. This is generally difficult because of several additional factors, such as Web caching, the existence of proxy servers and the stateless service model of the HTTP protocol. Several heuristics exist to address these problems. By definition, the heuristics yield inexact and variable results. It is, therefore, crucial to analyse and understand how good a particular heuristic is likely to be in a given environment. This paper reports on an investigation into the performance of a composite heuristic based on three published heuristics found in literature to identify sessions from the Web logs. We use the logs of a university Web server that records user ids for administrative reasons, which allows us to evaluate the heuristics against the concrete knowledge of user sessions. Consequently, the paper also proposes a strategy for future log analyses and makes recommendations for further work.

Keywords: heuristics, user tracking, user identification, session identification, log analysis, web usage mining.

1. Introduction

Web Usage Mining (WUM), a discovery of interesting user access patterns from Web server logs, has become the subject of intensive research, because

of its potential for personalized services, adaptive Web sites, target marketing in e-commerce, and organization and presentation of Web sites. With the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and Intranet technologies, to track and analyse user access patterns hidden in their Web server logs.

A Web server log is an important source for performing WUM because it explicitly records the browsing behaviour of site visitors. It provides details about file requests to a Web server and the server response to those requests. However, because of considerations of privacy, the logs, by default, do not record user ids.

For meaningful WUM, on the other hand, these requests must be identified into user sessions as semantic units of analysis. The difficulty of identifying users and user sessions from Web server logs has been addressed in by several researchers [1, 2, 3]. A solution to this problem is to create heuristics that capture in a logical way the behaviour of users and map it onto the Web logs.

The researchers have also examined the performance of the proposed heuristics[3, 4, 5]. This paper is a contribution to this analysis, based on the logs of a University Web site. These server logs, for operational reasons, maintain explicit user ids. The server under consideration also serves a fully known set of users, viz. students. The conditions under which the students work are also known. We can, therefore, analyse the logs to know exactly which student worked on the server at what time and for how long. Hence, by correlating this analysis with that thrown up by heuristics we are able to arrive at how well they perform. While these results are site-specific and hence difficult to generalise, the analysis enables us to formulate and propose a strategy for future analyses under different conditions and

improve one's understanding of user behaviour of a particular site.

The rest of this paper is organized as follows. Section 2 describes the logged data in general and the difficulties in using such data to analyse the user behaviour. Section 3 is a brief survey of the available heuristics and related work. Section 4 describes the heuristics combined for the evaluation. Section 5 explains the methodology used for the evaluation of heuristics. In section 6, we compare the performance of the heuristics for the site under consideration. Section 7 proposes a strategy for further work in evaluating heuristics and concludes the paper.

2. Web Server Logs

Web server log files are the primary source of data in which the activities of Web users are captured. These log files can be stored in various formats such as Common Log Format (CLF) or Extended Common Log Format (ECLF) as recommended by W3C[6], Microsoft and NCSA. An ECLF file is a variant of the CLF file simply adding two additional fields to the end of the line, the referrer and user agent fields. Each entry in the log file stored in ECLF describes the source of a request, the file requested, the date and time of the request, the URL referring to the requested file (referrer), the client environment (user agent), and other data such as server return code and the number of bytes transferred.

User requests for one URL frequently result in multiple entries in the server logs, independent of one another, representing requests to the server for each of the hyperlinked elements, such as images, style sheets and so on. The number of requests per day of a medium-large Web server is in the order of millions and a popular Web site can see its Web log growing by hundreds of megabytes every day.

The stateless service model of the HTTP protocol does not allow support for establishing long-term connections between the Web server and the user. The lack of explicit user identification in the log means that even the multiple requests generated by a single click cannot be assigned to the individual user who has initiated it with 100% certainty.

The process of user identification is mostly based on the IP address of the client machine that made the requests. This IP address may be of a machine used by only one user at a given time or it may be that of a proxy in which case it could represent a number of users whose requests are being routed through it. In the first case, the same machine may be used by different users over time, a fact that is impossible to deduce from the log data. In the case of a proxy, numerous requests to the Web server from users

connected to the proxy can occur simultaneously. In both cases, the tasks of inferring from the log data the individual users and the 'paths', i.e. the hyperlinks, traversed by each of them become non-trivial. There is another impediment to this process of user identification, viz. caching. There are various levels of caching embedded in the Web, mainly to expedite a user's access to the frequently used pages. Those pages requested by hitting the "back" button available in most browsers, (heavily used by the Web users [7]), are all retrieved from the Web browser cache. Also, proxy servers provide an intermediate level of caching at the enterprise level. The server log data cannot capture these cache hits, rendering it an incomplete source of user behaviour.

3. Related Work

Researchers have proposed various methods to resolve the problem of tracking users and their activities from the server log data and also highlighted their drawbacks.

Client-side data collection can be implemented by modifying the source code of an existing browser to enhance its data collection capabilities. A modified browser is much more versatile and will allow data collection about a single user over multiple Web sites. In [8], XMOsaic 2.6 was modified to record a user's browsing activity. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities.

Client-side data collection can also be implemented by using a remote agent. A remote agent developed as Java applet was introduced in [9, 10] as a client-side Web usage data acquisition system. When a user first enters a Web site, the remote agent is uploaded into the browser at the client side. Thereafter, it captures all required features of user interactions with the Web site and transfers the acquired data to a data acquisition server, called the acquirer. When the agent is uploaded to browser, it receives a globally unique session ID from the acquirer and labels all captured data sent to the server with that ID. In addition, the remote agent reports visiting of cached pages (whether at the proxy or at the browser) to the acquirer server which results in more accurate tracking as compared with what records at Web server logs. Thus, the acquirer can transparently store data captured by different agents as separate semantic units, i.e., user sessions, without further requirement for user session re-identification. The main drawback with this implementation is that running the remote agent at the client side requires users cooperation in enabling Java at their browsers.

The data mining results from the 1996 Olympics site [11] were obtained using cookies to identify site users and user sessions. An HTTP cookie issued by the server to the client browser identified a visitor to the Olympic Site. Since the server logged the cookie of every request, the requests coming from any given browser could be positively identified. The major downside of this method is that many users often choose to disable the browser features that enable the acceptance of cookies. It is also impossible to know whether more than one person visits the Web site using the same instance of a browser.

An innovative method, called page conversion was introduced in [12]. This mechanism involves software downloading and works as follows. First, a Web page is encoded into cipher by a server-side enciphering module. The original Web page is replaced by this enciphered Web page. Then, a client-side program, called deciphering module, deciphers these encoded data and displays the content to Web user. The deciphering module also reports the user behaviour to the Access Pattern Collection Server (APCS) before the data is deciphered and shown. By having the enciphering and deciphering mechanism, one can ensure that these Web pages will not be shown unless the deciphering module is called and the APCS is informed, preventing deliberate bypassing of the data collection process. (Of course, the enciphering/deciphering mechanism can be removed if the system allows users to bypass the data collection process.). Each line of the APCS log consists of access time, user name, host name, host address and the URL of the Web page accessed. With user access patterns properly collected, the individual Web user behaviour can be better captured and analysed by the corresponding data mining techniques. The violation of user privacy is the main concern in this method.

Lamprey [13], a tool for doing quantitative and qualitative analysis of Web-based user interfaces, tracks users by rerouting all of their Web navigation through a central tracking gateway. The central mechanism of Lamprey's user tracking system is the parsing of HTML pages and embedding of tracking information in every hypertext link in the page. When a user being tracked by Lamprey requests a page, the system fetches it and changes every URL in that page to reroute it through Lamprey. An altered URL includes all the parameters necessary for Lamprey to fetch the original page and return it to the user. Once the user sends a URL to the Lamprey application, all subsequent activities are logged in Lamprey log files.

A technique of dynamic page rewriting is used in [14, 15, 16] to track the user. In this method, when the user first submits a request to the Web site, the server returns the requested page rewritten to include a

hidden field with a session-specific ID. Each subsequent request of the user to the server, will supply this ID to the server, thus enabling the server to maintain the user's navigation history. An identifier timeout mechanism was also used to make sure different sessions from the same client are given different identifiers. This session-tracking method does not require any information on the client side and can therefore be always employed, independently of any user-defined browser settings. But this method restricts intermediate caching and does not correctly handle the exchange of URLs between people.

The heuristics proposed in [1, 3] can be used to help identify user sessions with relative accuracy in the absence of additional information such as cookies, user id or session id. The first heuristic states that two accesses from the same host but with different browser versions or operating systems are initiated from different visitors. The second heuristic states that if a web page is requested and this page is not reachable from previously visited pages, then the request should be attributed to a different user. The third heuristic is that all requests from the same host, browser and operating system within a threshold (usually 30 minutes) are considered to be part of the same session. The fourth heuristic says that the time spent on a page must not exceed a threshold (usually 15 minutes) to be part of the same session.

4. Selection of Session Identification Heuristics for Evaluation

In this study, we evaluate the performance of a composite heuristics based on three published heuristics. These heuristics are mainly to overcome the problems of proxies as well as to identify multiple users with the same, genuine IP address, as, for example, happens in a lab. The description of the three heuristics is as follows.

IP/Agent: Each different user-agent type for an IP address represents a different user.

A user-agent is an arbitrary assigned string, which shows a change in use of browser or operating system. The rationale behind this rule is that a user rarely employs more than one browser when navigating in the Web.

Referrer: A referrer is the URL of the page the client was on before requesting the current page. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, then the request should be attributed to a different user even if

the IP/Agent string is the same for the two consecutive page requests in the log.

The rationale behind this heuristic is that users generally follow links to reach a page and very rarely type URLs and use bookmarks.

Timeout: For the same combination of IP/Agent, if the time between two consecutive page requests exceeds a certain limit (15 minutes), it is assumed that the user is starting a new session.

The motivation behind this heuristic is that for logs that span long periods of time, it is very likely that users will visit the Web site more than once. Visitors who do not request pages within a certain time limit are assumed to have left the site and started new session.

The third and the fourth heuristics mentioned in the last paragraph of section 3 concerns about the timeout of a session and a page respectively. We combine the fourth heuristic with the first two in our evaluation

5. Methodology for Evaluating Heuristics

5.1 Environment

The Web logs used in this investigation came from the server for a student lab used exclusively to teach two Internet-related subjects. The students have to create a Web site each and then learn scripting for both client-side and server-side processing, including database connectivity. Each student is given an id and Web space. The server runs Microsoft Internet Information server 5.0 (IIS 5.0) on Windows 2000 advanced server platform. The lab has 20 workstations, each with a unique, hard-wired IP address. These machines primarily run Windows 2000 professional. Students access the server from the special lab, other labs or from outside, using either the university dial-up lines or some ISP. The university routes traffic from ISPs through two proxies. The traffic from on-campus computers and through university dial-up lines is not affected by any proxies. The university semester runs for 16 weeks during which time the students typically complete two assignments, some quizzes and a mini-project each. The lab has been running for almost five years. We chose the latest semester, viz. March-June 2003. Approximately 500 students enrolled in the two subjects.

5.2 Selection of logs

The server logs are created daily, always starting at midnight. We restricted this analysis to a total of 24 days, 10 in April 2003, seven each in May and June. The number of 'hits' recorded by the log ranged from over 4000 to more than 250,000 in a day. The seven days in May and June were deliberately chosen to cover three days before and after the deadline of the assignment that was due on the deadline, when we expected an increasing trend followed by a decline in the number of server hits.

5.3 Data Cleaning

Prior to evaluating the performance of heuristics to identify the users and their activities recorded on the server logs under the current Web log mechanism, data cleaning process should be done to filter out irrelevant log entries. In most cases, only the log entry of the HTML file request is relevant and should be kept for the user sessions. User requests for one URL frequently result in multiple entries in the server logs, independent of one another, representing requests for the hyperlinked elements, such as images, style sheets and so on. Since the main intention of the Web Usage Mining is to get a picture of the user's behaviour, it does not make sense to process such file requests. This also reduces the size of the data to be analysed.

Like most Web log analysis tools, the cleaning process employed in our method performs the following tasks. First, requests for non-HTML URLs are filtered out. Irrelevant items could be identified based on the suffix of the URL name in the log file. For instance, all log entries with filename suffixes such as GIF, JPG, JPEG, gif, jpg, jpeg are ignored. The set of suffixes could be adjusted as needed for particular Web sites by making changes to cleaning criteria. Next, other known useless data is filtered out like entries with particular server response status code such as "401", "403", "404" and "500" which means some error occurred in client or server side. All the entries, which record user id as unknown ("-"), are also removed as wrong ids. Finally, any extra spurious links such as mistyped URLs and spurious agents are removed.

5.4 Terminology

The following terms are used in evaluating the heuristics.

Hits: the number of individual requests to the server. These include, as explained above, requests for not only HTML documents, but also for gif, jpeg etc.

Views: the number of requests to the server after data cleaning is carried out, explained above. This ‘cleaned’ data is the base to which the heuristic is applied. The total number of views are also analysed using the explicit user ids recorded in the logs.

Sessions: A session is a sequence of page accesses performed by the user to accomplish a task. Sessions are defined on the basis of the amount of time spent on a single page.

Non-HTML hits: As explained above, these arise from a user’s request for an HTML page and are superfluous for the present purpose.

Errors: the number of hits that generated error messages from the server.

Spurious hits: the number of hits that did not contain any meaningful URL and/or Agent.

6. Evaluation

6.1 Data Cleaning

Table 1 gives the detail of the effect of data cleaning mentioned before and converts the number of Hits to Views. It covers the logs from 10 days in April. Data cleaning was carried out on the remaining logs as well with a similar pattern of conversion.

Date	Hits	Non-HTML Hits	Error Hits	Spurious agent or Url	Wrong id	Views
1/4	44264	12824	1298	823	325	28994
2/4	97333	24523	3730	1693	557	66830
3/4	31674	8591	1237	426	265	21155
4/4	13384	4470	591	381	138	7804
5/4	5683	1423	177	26	110	3947
6/4	9598	2773	265	61	129	6370
7/4	8884	2147	328	130	264	6015
8/4	13518	3669	845	78	336	8590
9/4	15942	3216	795	114	593	11224
10/4	5448	1012	347	55	211	3823

Table 1 - Data Cleaning: Converting ‘Hits’ to Views’

It is worth mentioning here that, for the purposes of this paper, we are interested in analysing the number of Views. However, the difference between the Hits and Views is still a demand on the server and could affect the server performance. Further, a

greater number of graphic and other images are also indicative of the type of site(s) under scrutiny. It is also a moot point if the number of ‘errors’ is indicative of the user performance that ought to be taken into account in personalising a Web site.

6.2 User Sessions

Tables 2 to 4 give the details of the performance of the heuristic for the three periods under consideration. The number of Views arrived at after the Data Cleaning step is converted to User Sessions based on the heuristic and also the explicit identification of students in the logs.

The composite heuristic used here works in the following way. First, we use IP/A and Referrer to identify session and then we apply the page timeout of 15 minutes. The number of distinct user sessions is also calculated separately with the help of usr ids. In this case, two separate sessions of the same user are tracked by IP address and Timeout heuristic.

6.3 Analysis

Several points emerge from this analysis. First, the heuristics have over-estimated the number of user sessions on all days. This is to be expected because the numbers of sessions identified through user ids represent near-complete knowledge of the local situation. Heuristics are more generic. The exact number of user ids found in the log of any particular day is the least number of sessions possible. However, an interesting phenomenon came to our notice when we examined the raw (cleaned) data further, viz. that the numbers of user sessions in both cases get inflated for two reasons. The first reason is to do with the use of proxies. The university maintains two proxies to balance the workload of the main servers, which come into operation whenever a student uses an ISP. It can happen that two consecutive accesses from the same student come to the server through different proxies within the time-out threshold of 15 minutes because of load balancing algorithm. The second factor in inflating the numbers is the way the logs are maintained. As mentioned before, each day’s log starts at midnight and goes on for the next 24 hours. The logs show that there are a good number of students who work around mid-night. Consequently, when a student starts his/her session before midnight and continues for some time after midnight, that session is split into two logs and is counted twice. To check the extent of such double counting, we merged log files for several days together and ran the analysis again. Tables 5 and 6 illustrate the results. For the 10-day period in April 2003, the number of sessions

split across midnights comes to 35. This may not amount to much in the current study but it is indicative of what can happen to the logs, depending upon the policies followed at a given site.

The second point to emerge from this analysis is that the range of differences between the estimates of user sessions by the composite heuristics and those by the user ids is rather large, from just over 8% to more than 100%. This makes it difficult to judge the relative performance at this stage, beyond saying that the composite heuristic ‘over-estimate’ these numbers. Ideally, one would like to estimate the scale of variation so that the heuristics could be used with more confidence. Further investigation is needed to arrive at more quantifiable understanding..

The third point is about the student behaviour in submitting their assignments, as exhibited by the number of hits and sessions. The submission dates of two assignments were 7th of May and 4th of June and their effect cannot only be anticipated qualitatively but also quantitatively. Tables 3 and 4 reflect these patterns. This has implications for the server and network administrators, the students and the academics in charge. Peak loads can be identified in advance and students can be shown the results of such analyses to help them to submit their assignments relatively trouble-free.

Date	Views	Sessions (heuristics)	Sessions (user ids)	Diff-erence	%Diff
1/4	28994	370	342	28	8.19
2/4	66830	769	654	115	17.58
3/4	21155	316	273	43	15.75
4/4	7804	146	117	29	24.79
5/4	3947	97	74	23	31.08
6/4	6370	117	102	15	14.71
7/4	6015	177	150	27	18.00
8/4	8590	182	167	15	8.98
9/4	11224	274	251	23	9.16
10/4	3823	133	103	30	29.13

Table 2: Performance of Heuristics (April 2003)

Date	Views	Sessions (heuristics)	Sessions (user ids)	Diff-erence	%Diff
4/5	7348	169	133	36	27.07
5/5	9446	353	314	39	12.42
6/5	25896	1370	680	690	101.47
7/5	41663	1399	1012	387	38.24
8/5	9011	297	240	57	23.75
9/5	3315	152	122	30	24.59
10/5	3259	96	77	19	24.68

Table 3: Performance of Heuristics (May 2003)

Date	Views	Sessions (heuristics)	Sessions (user ids)	Difference	%Diff
1/6	53482	857	677	180	26.59
2/6	75853	1514	1228	286	23.29
3/6	207874	2746	2108	638	30.27
4/6	105283	1749	1187	562	47.35
5/6	29786	617	414	203	49.03
6/6	14470	445	289	156	53.98
7/6	7738	255	190	65	34.21

Table 4: Performance of Heuristics (June 2003)

Date (from)	Date (to)	Views	Cummulative sessions (individual run)	Sessions (batch run)	Diff-erence
1/4	2/4	95824	1139	1123	16
1/4	3/4	116979	1455	1429	26
1/4	4/4	124783	1601	1572	29
1/4	5/4	128730	1698	1667	31
1/4	6/4	135100	1815	1784	31
1/4	7/4	141115	1992	1960	32
1/4	8/4	149705	2174	2140	34
1/4	9/4	160929	2448	2410	38
1/4	10/4	164752	2581	2539	42

Table 5: Number of sessions spanned over logs based on the heuristic (April 2003)

Date (from)	Date (to)	Views	Cummulative sessions (individual run)	Sessions (batch run)	Diff-erence
1/4	2/4	95824	996	985	11
1/4	3/4	116979	1269	1250	19
1/4	4/4	124783	1386	1364	22
1/4	5/4	128730	1460	1436	24
1/4	6/4	135100	1562	1538	24
1/4	7/4	141115	1712	1687	25
1/4	8/4	149705	1879	1852	27
1/4	9/4	160929	2130	2099	31
1/4	10/4	164752	2233	2198	35

Table 6: Number of sessions spanned over logs based on user id (April 2003)

7. Conclusions and Recommendation for a Strategy to Evaluate Heuristics

This paper has reported on the investigation into the efficiency of heuristics that may be used to identify user sessions, based on the analysis of Web logs. This was done on the basis of knowing the ‘local’ circumstances, policies and procedures, which allowed for much better estimates of user sessions. The heuristics, by their very nature, are generic, without this local knowledge. It is still too early to draw definite quantitative conclusions about the efficiency of these heuristics in combination or individually. However, our investigation allows us to

recommend a strategy for future work, as outlined below.

1. Start with the log data bearing in mind that it is just the raw data and needs to be 'cleaned up'.
2. Formulate and use a 'Cleaning' procedure. The cleaning procedure will depend upon the local circumstances and policies. Thus, for example, the logs we analysed contained error messages, spurious information, non-HTML requests and wrong (user) ids. It is possible to use more than one log file to log these entries separately. Also, the logs may be hourly, daily, weekly or any time period as determined by the Web administrators. The cleaning procedures must incorporate such knowledge.
3. If any part of the logs contains explicit user ids or session ids, use them to isolate the subsets of data where such local practices will make it easier to understand how much the heuristics vary in their analysis.
4. Analyse the remaining, cleaned data that does not contain any user/session ids, on the basis of the heuristics. Use the understanding from step 3 to refine the analysis, as necessary.

The fourth step in effect use the understanding derived from the third step in an empirical way. That is to say that the user behaviour as understood by carrying out step 3 is expected to remain unchanged.

There is more detailed analysis under way. First of all, we plan to test the statistical significance of each heuristic. Then, various combination of heuristics will be similarly analysed. We have also come across some anomalous results, including the 'outlier' of 101.47% more user sessions on the basis of heuristics (see Table 3). These and more detailed analysis of individual heuristics applied in different situations would be the tasks for the future.

References

1. Pirolli, P., Pitkow, J.E., and Rao, R., *Silk From a Sow's Ear: Extracting Usable Structure from the World Wide Web*. "Conference on Human Factors in Computing Systems (CHI 96)". Vancouver British Columbia, Canada 1996
2. Pitkow, J.E., *InSearch of Reliable Usage Data on the WWW*. "The sixth International World Wide Web Conference". Santa Clara, California 1997
3. Cooley R., Mobasher B., and Srivastava J., *Data preparation for mining world wide web browsing patterns*. "Journal of Knowledge and Information Systems". **1**(1): p. 5-32. Springer-Verlag February, 1999
4. Berendt, B., Mobasher B., Spiliopoulou, M., and Wiltshire, J., *Measuring the accuracy of sessionizers for web usage analysis*. "Proceeding of the Workshop on Web Mining, First SIAM International Conference on Data Mining": p. 7-14. Chicago, IL 2001
5. Berendt, B., Mobasher B., Spiliopoulou, M., and Nakagawa, M., *A framework for the evaluation of session reconstruction heuristics in web usage analysis*. "INFORMS Journal of Computing". **15**(2). 2003
6. www.w3.org/Daemon/User/Config/Logging.html. last accessed: 22nd of September, 2003,
7. Greenberg, S. and Cockburn, A., *Getting Back to Back:: Alternate Behaviors for a Web Browser's Back Button*. "Proceeding of the 5th Annual Human Factors and Web Conference". NIST, Gaithersburg, Maryland, USA June 3rd, 1999
8. Tauscher, L. and Greenberg, S., *Revisitation patters in World Wide Web navigation*. "In ACM SIGCHI '97 Proceedings of the Conference on Human Factors in Computing Systems": p. 22-27. ACM Press Atlanta, Georgia, USA March, 1997
9. Shahabi C., Zarkesh A., Adibi J., and V., S., *Knowledge Discovery from Users Web-page Navigation*. "Proceedings of the IEEE RIDE97 Workshop". April, 1997
10. Shahabi C., Banaei-Kashani F., and J., F., *A Reliable, Efficient, and Scalable System for Web Usage Data Acquisition*. "WebKDD'01 Workshop in conjunction with the ACM-SIGKDD". San Francisco, CA August, 2001
11. Elo-Dean, S. and Viveros, M., *Data mining the IBM official 1996 Olympics Web site*. "Technical report". IBM TJ Watson Research Center 1997
12. I-Yuan Lin, X.-M.H., Ming-Syan Chen, *Capturing User Access Patterns in the Web for Data Mining*. "11th IEEE International Conference on Tools with Artificial Intelligence": p. 345. Chicago, Illinois November 08-10, 1999
13. Felciano, R.M. and Altman, R.B., *Lamprey: Tracking Users on the World Wide Web*. "AMIA Annual Fall Symposium". Hanley & Belfus. Washington, D.C. 1996
14. Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Dayal, U., *From User Access Pattern to Dynamic Hypertext Linking*. "Fifth International World Wide Web Conference". Paris, France May 6-10, 1996
15. El-Ramly, M. and Strulia, E., *Web-usage Mining and Run-time URL Recommendation for Focused Web Sites: A Case Study*. "Journal of Software Maintenance and Evolution: Research and Practice". **00**: p. 1-7. John Wiley & Son, Ltd 2000
16. Nan Niu, E.S., Mohammad El-Ramly,, *Understanding Web Usage for Effective Dynamic Web-Site Adaptation*. "4th International Workshop on Web Site Evolution". Montréal, Canada October, 02, 2002



KANSEI Information Processing for Perception of Vowel Sounds

Chandrajith Ashuboda Marasinghe,¹ Ajith P. Madurapperuama,² Stephen G. Lambacher,¹
William L. Martens,³ Michael Cohen,¹ and Susantha Herath⁴

¹University of Aizu, Aizu-Wakamatsu, Fukushima-ken 965-8580, Japan.

²Faculty of Information Technology, University of Moratuwa, Colombo 08, Sri Lanka.

³McGill University, 555 Sherbrooke Street West, Montreal, Quebec, Canada.

⁴St. Cloud State University, St. Cloud, Minnesota 56301-4498, U.S.A.

email : d8032103@u-aizu.ac.jp

Abstract

In this paper we investigate the specifications of the KANSEI information processing in perception of American English vowel sounds. A common perceptual space for 10 American English vowel sounds was derived for two groups of listeners, a group of native speakers of the Japanese language, and a group of native speakers of Sinhala, a language of Sri Lanka. The stimuli used in the experiment were the ten vowel sounds synthesized using the often utilized formant frequency values published by Peterson and Barney in 1952. Subsets of these two groups made ratings on 12 KANSEI bipolar adjective scales for the same set of sounds, each of the two groups using anchoring adjectives taken from their native language. Though there was no evidence of any difference between the two groups' in their INDSCAL-derived perceptual dimensions for these vowel sounds, the adjectives were used differently in describing those same perceptual dimensions by the two groups. The results of semantic differential analysis (SDA) support the conclusion the two groups' ratings on 12 KANSEI bipolar adjective scales related somewhat differently to the dimensions of their shared perceptual space. Though a few of the adjectives were used to describe similar perceptual variations, one implication of this investigation is that caution be exercised in generalizing semantic differential ratings obtained in one language, especially when those ratings are intended to aid in the interpretation of data from listeners speaking a different native language.

Keywords: KANSEI information processing, perceptual space, semantic differential analysis.

1. Introduction

Although the last 50 years have witnessed a rapid growth in the understanding of vowel articulation and acoustics, most contemporary theories of speech perception have concentrated on vowel perception [1] [2]. There are at least as many meanings of "meaning" as there are disciplines which deal with perception of vowel sounds, and of course, many more than this because exponents within disciplines do not always agree with one another. Nevertheless, definitions do tend to correspond more or less with the purposes and techniques of the individual doing the defining, focusing on that aspect of the phenomenon which his discipline equips him to handle. Thus, the sociologist or anthropologist typically defines the meaning of a sound in terms of the common features of the situations in which it is used and of the activities which it produces [3].

Of all the implications that inhabit the nervous system that "Little black box" in psychological theorizing—the one we call "perception" is held by common consent to be the most elusive. Yet, again by common consent among social scientists, this variable is one of the most important determinants of human behavior. It therefore behooves us to try, at least, to find some kind of objective index. To measure anything that goes on within "the little black box" it is necessary to use some observable output from it as an index. To the problem yet another way, we wish to find a kind of measurable activity or behavior of sign-using organisms which is maximally dependent upon and sensitive to meaningful states, and minimally dependent upon perception. In the search for such indices of perception, KANSEI Information processing can help to find observable output for perception of vowel sound.

2. *KANSEI* Information Processing (KIP)

KANSEI Information Processing (KIP) is a new branch of Information Processing Technology born in Japan [4] [5] [6]. *KANSEI* is a Japanese word where the syllable *kan* means sensitivity and *sei* means sensibility, that does not have a direct counterpart in Western languages, or, however, every attempt to translation captures just some of the aspects of *KANSEI*. In 1997, the *KANSEI* evaluation special project started as a five-year interdisciplinary project at University of Tsukuba in Japan. Because it was found that the term '*KANSEI*' was used in different meanings by different participating researchers, an initial study mapped there meanings[5]. The researchers in the project were asked to give their definition of *KANSEI*. These statements were analyzed, and key words were clustered to five main aspects as follows.

1. *KANSEI* is a subjective effect which cannot be described by words alone.
2. *KANSEI* is a cognitive concept, influenced by a person's knowledge, experience, and character.
3. *KANSEI* is a mutual interaction between the intuition and intellectual activity.
4. *KANSEI* entails a sensitivity to aspects such as beauty or pleasure.
5. *KANSEI* is an effect for creating the images often accompanied by the human mind.

Most of all, it's important to understand *Kansei* implies that human behaviors can change dynamically, and indicates flexible and dynamic approaches are needed in the various fields of study. The aim of *KANSEI* study is to seek the structure of emotions which exists beneath human behaviors [7][8]. This structure is referred to as a person's *KANSEI*. In the art and design field, *KANSEI* is one of the most important elements which brings the willing or power of creation. In research by Harada, it was found that the attitude of a person in front of art work and design is not based on logic but on *KANSEI* [5]. *KANSEI* is an ability that allows humans to solve problems and process information in a faster and personal way. In every action performed by a human being, traces of his/her *KANSEI* can be noticed, as well as his/her way of thinking, of solving problems, of his/her personality. Therefore, *KANSEI* is not a synonym for emotion, although it can be related to emotion, but refers to the human ability of processing information in ways not just logical.

2.1 KIP for vowel perception

The concept of *KANSEI* is strongly tied to the concept of perception and sensibility. *KANSEI* is an ability that allows humans to solve problems and process information in a faster way. A movement of a human's nerves system receives big influence in the external world information detected by the senses, and it appears as a sensitivity reaction. *KANSEI* information is subjective and ambiguous, and it depends strongly upon an individual or a situation[9]. Therefore, it differs from the logical knowledge information, which has been the object of usual information processing. This vowel perception project is based upon the idea that the way in which words are used can be quantified via Semantic Differential Analysis (SDA), a well established method for the measurement of word meanings using bipolar scales with adjectives of opposite meaning anchoring either end of each scale [3]. Indeed, the application of SDA in investigations of the human description of sound events has a history dating back to the 1950's [2]. It is broadly used to arrange and unify a feeling or a sense which was expressed with language and to evaluate it. Sound is described by many terms. The language is subjective and ambiguous. However, to describe sound with language requires some descriptive evaluations of sound. The relation between humans and sound is very convenient if the character of sound can be evaluated using adjectives. Perhaps a function modelling human taste could be developed if one could analyze how a human would feel listening to a vowel sound. The primary motivation for this study is to evaluate the semantic properties of sound perception using *KANSEI* adjectives to describe vowel sounds. Another motivation for this study was the refinement of adjective scales for use in subsequent studies of perception using groups of listeners with various native languages.

Two listening experiments were executed to establish both perceptual and semantic scale values for each of 10 American English vowel sounds. In the first experiment, Japanese and Sinhalese listener's gave dissimilarity ratings for all pairwise comparisons of the vowel sound stimuli. Submitting these obtained dissimilarity ratings to **INDividual Differences SCALing (INDSCAL)** analysis yields a perceptual space for the stimuli. In the second experiment, the Japanese and Sinhalese listeners made ratings on a set of 12 *KANSEI* bipolar adjective scales for each of the 10 stimuli. It was not assumed at the outset that ratings on these 12 adjective scales should necessarily capture the most salient differences between the stimuli. These 2nd experiment data were analyzed through **Principle Component Analysis (PCA)**. It was hoped that the results would provide a better understanding of how

adjectives from each language are used by listeners to describe the sound of American English Vowels as part of an ongoing research project.

2.2 Individual Differences SCALing (INDSCAL).

The purpose of *INDSCAL* is to represent objects, whose dissimilarities are given, as points in a metrical space. The distances in the space should be in accordance with the dissimilarities as well as is possible. Besides the configuration a Salience matrix is calculated. Individual subjects may differ in how they form judgments of global dissimilarity, and so a refined method for doing a weighted *INDSCAL* analysis [10] [11] that takes such individual differences into account is to be recommended. Use of *INDSCAL* analysis as a powerful means for deriving an interpretable representation of the dimensions underlying reported inter-stimulus dissimilarities obtained from a potentially inhomogeneous group of subjects, each of which may place different weights upon each of the perceptual dimensions.

2.3 Principal Components Analysis (PCA)

PCA is well known statistical method used for essence of information from vast amount of data obtained from subjective evaluation experiments [12][13] [14]. Subjective evaluation results contain somewhat redundant data, and *PCA* is useful in order to reduce the driving forces governing the redundancy. *PCA* is a method of transforming a number of variables into one or a few linearly independent representative variables (principal components) with the least amount of information lost. Each principal component is a linear combination of the original variables, and all principle components are orthogonal to each other. The meaning of *PCA* is often confused with Factor Analysis (FA) because they share the same usage of replacing a large set of observed variables with a smaller set of new variables. However, *PCA* is used for reducing a number of variables to explain the overview of observed data, while FA is used for discovering a set of latent (unobservable) variables underlying the subjective judgments in adjective ratings.

3. Methods

3.1. Stimuli

Ten American English vowel sounds were digitally synthesized [Table 1]. Similar sounds have been used in

speech science since Peterson and Barney first published these vowel formant data in 1952 [15].

3.2. Subjects

46 Japanese native speakers were recruited for Dissimilarity Rating Task and 52 Japanese native speakers were recruited for Semantic Differential Rating Task. All subjects were undergraduates at the University of Aizu in Fukushima Prefecture Japan. 40 Sinhala native speakers were recruited for both the dissimilarity and semantic differential rating task. All subjects were undergraduates at the University of Colombo Sri Lanka.

3.3. Listening Tasks

The two listening experiments were completed for each listener. These two listening tasks were described to each listener just before each task was executed. All instructions are given by listeners native language.

3.3.1. Dissimilarity Rating Task.

Listeners were asked to give global dissimilarity ratings on a 5-point scale for 90 pairwise comparisons of the American English vowel sound stimuli. The instructions were to listen to each stimulus pair once, and the rate their global dissimilarity without respect to any particular property. A response of "1" implied that the two samples were perceived as "almost exactly the same" and a response of "5" implied that the two stimuli were perceived as "almost completely different." Each pair of stimuli was presented twice, in a random order, always with a 1-second inter-stimulus interval and a 5-second inter-trial interval (the time between each presented pair of stimuli in which responses were recorded).

3.3.2. Semantic Differential Rating Task

Listeners were asked to rate each stimulus on 12 *KANSEI* bipolar adjective scales. Again, a 5-point scale was employed, but in this case, a response of "1" indicated that the vowel stimulus was best characterized by the adjective anchoring one end of the semantic differential scale, while a response of "5" indicated that the stimulus was best characterized by the adjective anchoring the other end of the scale. Figure 2 shows the 5-point, *KANSEI* bipolar adjective scale. Listeners were instructed to give a response of "3" if neither of the anchoring adjectives characterized the stimulus. The 10 stimuli were presented in a order for each of 12 adjective pairs, with a 5-second

IPA symbol	ASCII code	Example word	Vocal Articulation		
			lips	tongue	mouth
i	EE	bead	unrounded	front	closed
I	IH	bid	unrounded	front	near-closed
E	EH	bed	unrounded	mid-front	open
/	AE	bad	unrounded	front	near-open
A	AH	bod(y)	unrounded	back	open
U	UH	bud	unrounded	mid-back	open
O	AW	bawd	rounded	mid-back	open
U	OO	bud(dhist)	rounded	near-back	near-closed
u	UU	booed	rounded	back	closed
ā	ER	bird	rounded	back	closed

Table1: 10 Stimuli - Vocal articulation for production of 10 American English vowels (those investigated in Peterson and Barney, 1952) [15], including for each vowel its IPA symbol, a code of ASCII characters used for graphics, and an example word[16].

Japanese	English	Sinhala
hakkiri-shitha - kumotta	clear - unclear	pirisidu - apirisidu
surudoī - nibui	sharp - dull	thiunu - mota
akaurui - kurai	bright - dark	depthimath - aduru
takai - hikui	high - low	uase - pahathe
omoi - karui	heavy - light	bara - sallu
sunda - nigotta	transparent - muddy	vinividapenena - bora
konpakuto na - hirogatta	compact - diffuse	asirinu - pathirinu
ochitsuita - sozoshii	calm - clamorous	sanesune - tadagoshakari
nameraka na - arai	smooth - rough	sinidu - ralu
atsui - usui	thick - thin	gana - thuni
dodotoshita - hikaemena	magnificent - humble	vishisheta - ashisheta
hakuryokunoaeu - yowai	powerful - weak	shakthimath - dureweala

Table 2: Corresponding *KANSEI* bipolar adjective pairs in three languages, listed in order of presentation for subsequent semantic differential ratings. The Japanese adjectives were translated from the English by a native speaker of Japanese, and the Sinhalese adjectives were translated from the English by a native speaker of Sinhala.

of interval between each individual stimulus presentation. Adjective scales were given by listeners native language. Table 2 shows Corresponding bipolar adjective pairs in three languages, listed in order of presentation for subsequent semantic differential ratings.

The working assumption in the current study is that complex sounds having multiple perceptual attributes have a mental structure that can be quantitatively captured in terms of a multidimensional perceptual space that is distinct from the words that might be used to describe the individual perceptions occupying that space. It is hypothesized that the dimensions of perceptual space for a small set of stimuli may be common among

groups of listeners with differing native languages. It is further hypothesized that the words used by multilingual groups of listeners may share common underlying semantic structures when used to describe that small set of stimuli. Determining whether or not either of these hypotheses can be supported by experimental data is the primary goal of this study.

4. Results and Discussion

4.1. Dissimilarity Ratings Results

Ratings of dissimilarity were reported for all pairwise comparisons of the 10 American English vowel stimuli native speakers of Japanese and native speakers of Sinhala. Each stimulus pair was presented only once, and listeners used a 5-point scale for their ratings. A separate 10 X 10 matrix of dissimilarity data was constructed for each of the listeners, and these were combined into a single submission for *INDSCAL* analysis, using the *ALSCAL* routine found in the *SPSS* statistical analysis software.

One way to determine whether two groups of listeners share a single, common perceptual space is to examine the ways in which they generate dissimilarity judgments for pairwise comparison of stimuli taken from the set of vowels of interest. Figure 1 shows the *INDSCAL* derived common perceptual space common to the two groups of subjects. This result, of course, does not prove that the most salient perceptual dimensions are the same for the two groups of subjects; rather the result provides no evidence that the groups differ in their global responses to the stimuli. The conclusion that the null hypothesis should be retained is not the same as proving that no differences exist.

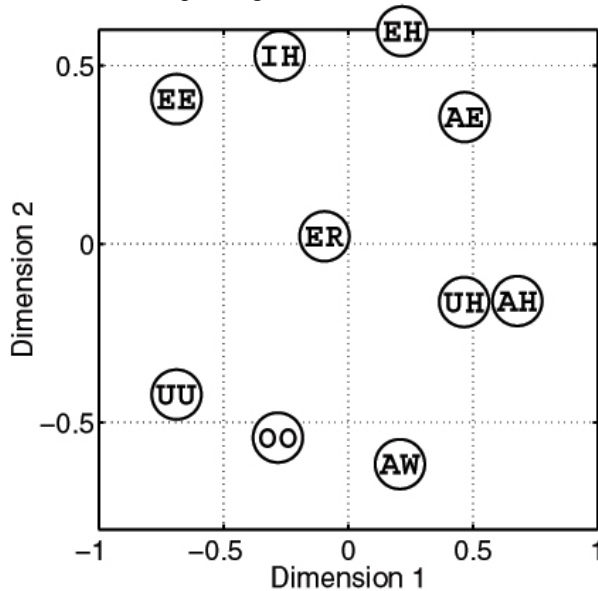


Figure 1: *INDSCAL* derived common perceptual space for Japanese and Sinhalese listeners.

4.2. Semantic Differential Rating Result

Ratings on the 12 *KANSEI* bipolar adjective scales shown in Table 2 were collected for the 10 vowel sound stimuli from the 52 Japanese and 40 Sinhalese listeners. A single 10 X 12 matrix of bipolar adjective ratings data was constructed for each listener, and these were combined into a single submission for Principal Components Analysis PCA, using the *FACTOR* routine found in the *SPSS* statistical analysis software.

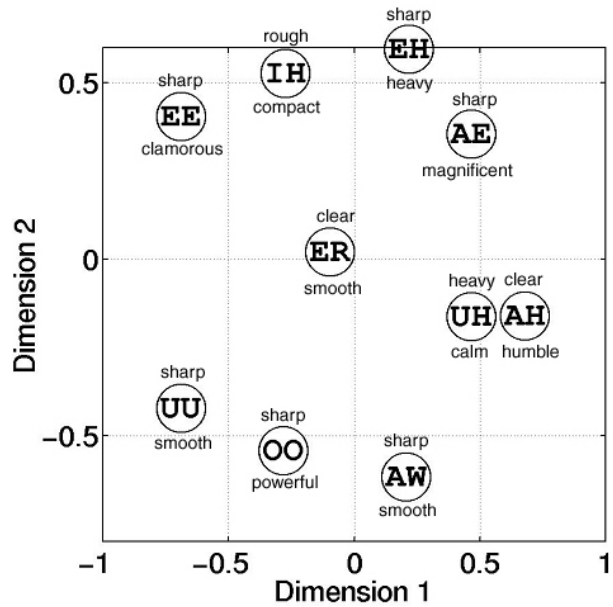


Figure 2: PCA derived *KANSEI*/representation of vowel sounds by native speakers of Japanese. 1st *KANSEI*/principle components are shown at the top of the vowels and 2nd *KANSEI*/principle components are shown at the bottom of the vowels.

Figures 2 & 3 show *KANSEI* representation of vowel sounds by native speakers of Japanese and native speakers of Sinhala.

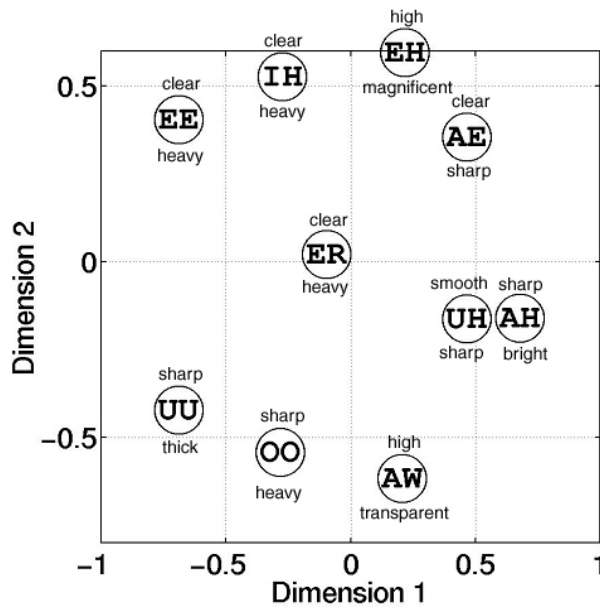


Figure 3: PCA derived *KANSEI*/representation of vowel sounds by native speakers of Sinhala. 1st *KANSEI*/principle components are shown by top of the vowels and 2nd *KANSEI*/Principle components are shown by bottom of the vowels.

5. Conclusion

Taken together with the results of the *INDSCAL* results and the *PCA* results support the conclusion that adjectives are used differently in describing the same perceptual dimensions by native speakers of the Japanese and native speakers of Sinhala. A single, common perceptual space for a American English Vowel sounds was derived for the two groups of listeners, who were judged to share similar global perceptual responses due to the overlap of the individual dimensional weights obtained from the *INDSCAL* analysis. Ratings on 12 *KANSEI* bipolar adjective scales for the same set of sounds showed that the Japanese and Sinhalese *KANSEI* representations are somewhat differently to the dimensions of their shared perceptual space. Sharpness is the most common *KANSEI* for Japanese and clearness is the most common *KANSEI* for Sinhalese. With regard to the potential for generalizing semantic differential ratings obtained in one language to aid in the interpretation of data from listeners speaking a different native language, the results of the current study suggest that caution be exercised. It was hoped that the results would provide a better understanding of how *KANSEI* bipolar adjectives from each language are used by listeners to describe the sound of American English Vowels as part of an ongoing research project.

References

1. D. Kewelely-Port and B.S. Atal, "Perceptual differences between vowels located in a limited phonetic space", *J. Acoust. Soc. Am.*, 1726--1740 (1989-85-4).
2. L. N. Solomon, "Semantic Approach to the Perception of Complex Sounds", *J. Acoust. Soc. Am.* Vol 30-5 (1958-3).
3. C. E. Osgood, G. J. Suci and P. H. Tannenbaum "The Measurement of Meaning", University of Illinois 1957.
4. http://www.ergolabs.com/kansei_engineering.htm
5. Harada A., "The framework of *Kansei* engineering", Report of Modelling the Evaluation Structure of *Kansei*, pp49-55, 1997.
6. A. Camurri, S. Hashimoto, M. Ricchetti, R. Trocca, K. Suzuki, G. Volpe, "EyesWeb - Toward Gesture and Affect Recognition in Dance/Music Interactive Systems." *Computer Music Journal*, MIT Press, 1999.
7. A. Camurri, "The Technology of Emotion", Proceedings of the AIMI International Workshop on *KANSEI* - 1997, University of Genoa and AIMI.
8. A. Camurri, S. Hashimoto, K. Suzuki, R. Trocca (1999). "Kansei analysis of dance performance." *Proc. Intl Conf IEEE Systems Man and Cybernetics 1999, Tokyo*.
9. Y. Watanabe, "Kansei Engineering for control of effects processing for Musical sound", Master's thesis, University of Aizu, March 1999.
10. I. Borg and P. Groenen, "Modern Multidimensional Scaling: Theory and Applications", Springer-Verlag, 1997.
11. R. N. Shepard, A. K. Romney, and S. B. Nerlove, "Individual differences and multidimensional scaling", Seminar Press, 1972.
12. D. J. Kister and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis", *J. Acoust. Soc. Am.* 1637-1647 (1991-3).
13. W. L. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proc. Int. Computer Music conf.*, Champaign-Urbana, IL, Sept. 1987.
14. G. J. Sandell and W. L. Martens. Perceptual evaluation of principal-components based synthesis of musical timbres. *J. Audio Eng. Am.*, 43, No 12: 1013-1028, 1995.
15. G. E. Peterson and H. L. Barney, "Contral Methods Used in a Study of the Vowels" *J. Acoust. Soc. Am.* 175--184 (1952-3).
16. P. Ladefoged, "Vowels and Consonants", Balackwell Publishers, 2001.



Chip-based National ID Card --Characteristics & Feasibility

Chih-Chun Chang
2033 K St. NW, Washington D.C., 20052, U.S.A.

Cyber Security Policy & Research Institute
Department of Computer Science, The George Washington University
ccc987@gwu.edu

Abstract

The main object of this paper is to explore the feasibility of chip-based National ID cards and related issues. The paper first listed the benefits of the chip-based ID card system listed in the paper, such as powerful calculation, complicate logic control, large storage abilities, security mechanism, and multiple functionality cost saving. Besides, some controversial issues are thoroughly discussed to clarify the cons of chip-based national ID cards, such as database problems, invasion of privacy and anonymity, information management and inside threat, technological challenges, crime issues, biometrics, and problems of multi-purposed card. In addition, the paper conducts a case study to explore the implementation status of chip-based national ID system in Asian countries and various challenges in the process. The last part of the paper includes some suggestions and potential solutions for issuing the chip-based national ID cards.

1. Introduction

Several countries have already endeavored to use chip-card technology to store personal information. A national ID card with holder's photo picture and other information can be used for identification. Besides this seeable information, the digitalized data of personal data also can be store in the magnetic strip on the card. The modern technology has allowed a chip like a small computer embedded on the card, which can carry more information. To improve the effectiveness of E-service and the efficiency of information management, the chip-based national ID system would become one of the tools towards the circumstance of e-government. However, some issues catch great public attention while the feasibility of the chip-based national ID card system is concerned. The possibilities that the system

might be hacked always exist, and there are potential risks of infringing the citizens' privacy and freedom.

National ID system is still a controversial issue in many countries. The purpose of National ID system could be an exercise of political or religious discrimination, a tool of social engineering, fighting crime, national registration, and identification [1]. The groups advocating human rights, privacy and liberty are opposed of the national ID system, whereas the groups focusing on the convenience of information management and the development of e-business tend to favor the national ID system.

Many countries in today's world use plastic cards to carry information, such as driver's license, school identification cards, and credit cards. Modern technology has allowed a chip to be placed inside a wallet-sized plastic card. Such a card is often referred to as a chip-based card. A chip-based card could be programmed to perform multiple tasks and store information. The chip-based National ID Card could be used to store information like personal identification, financial information, medical information, and e-commerce related information, such as a digital certificate and digital signature. Eventually the card could be used to store information on a person's fingerprints or even DNA.

The placement of the chip inside a traditional plastic card makes it tamper-resistant and portable. Smart Cards can be used to store digital money, pre-paid phone calls, or personal identification such as a driver license. A chip-based identification card could display information on the surface just like current Driver's license, and additional information can be stored on the chip. Using a card reader a person could retrieve the chip-based information.

Many countries have debated the issues of whether or not to have a National ID Card. The ideas of National ID system or cards focused on improving the security of whole society to solving problems, such as terrorism, and illegal immigrations [2]. It seems inevitable that the introduction and usage of chip-based

National ID Cards would reduce personal privacy while enhancing public security as well as normal National ID Cards. The prospect of chip-based National ID Cards deserves much discussion to determine what is best for both the individual and society.

2. Chip-based National ID Card

2.1 National ID Card

In general, most ID card systems have support registers containing information parallel to those on the cards. Regional or central authorities often keep the registers and an integrated system is usually established in favor of the basic function of government administration. For government agencies, the numbers of ID cards become registration numbers for identification. In those countries where ID cards are issued, the holders' basic information is printed on the cards, such as names, date of birth, and sex. Sometimes the holders' signatures and photographs are also included [3]. Even the information on the cards is believed to be authentic, the majority of the countries issuing ID cards usually verify the holders' legitimacy by checking out whether their faces match the photos on the cards or requiring the holders to present more information [4].

2.2 Chip-based National ID Card

With the advent of digital era, there are increasing countries storing identification information in plastic cards, rather than paper-based cards, because the cards with magnetic stripe or bar codes seem much more durable and can store digitalized data [5]. The digitalized personal data can be stored in the magnetic strip on the card. However, more and more fraud activities of the card magnetic strip are growing rapidly, including card altering and re-embossing, secret data skimming, and data re-encoding. These fraud activities may lead to huge economic loss, especially to credit card companies. Magnetic strip cards are truly weak in security [6].

With the technological improvement, a chip can be placed inside a wallet-sized plastic card. Such a card is often referred to as a chip-based card or Smart Card. Unlike magnetic strip cards, which only store few bytes of data, the Smart Cards can carry a hundred or more times of information than magnetic stripe cards. In addition, the chip-based cards are integrated circuit (IC) with a microprocessor or another crypto-processor, RAM, and ROM inside. The Smart Card actually is a powerful microcomputer. It can run complex algorithms like generate RSA key pairs and recognize

illegal signal and other security features. The new type of cards not only improves the efficiency of administration, but also provides a better way to protect the sensitive identification information.

There are many advantages of the Smart Card technology. Smart Card advocates described a Smart Card as an "active token." Although chip-based card has computing power, the packaging technology makes it exactly the same size as "credit card" and against breakage, rubbing, and bending. The chip-based card is viewed as the most portable and most secure storage [7].

How secure are the Smart Cards? Tracking the data flow or the sequence of commands based on the protocol between the card and terminal, using electron microscope, UV, or X-ray to inspect the internal structure of the mask, disturbing the fault or power supply are possible attacks to the chip-based card. However, to achieve these attacks, it needs to have physical possession of the card, intimate card knowledge of hardware and software, and special equipment. It costs a lot to either crack the code or do electrical observation; on the other hand, it is difficult for physical attacks because a Smart Card offers hardware protection [8].

Besides the basic technique such as photo picture, micro printing, or hologram to protect ID card, the assess right to the card is also need to be protected. The objective of the protection mechanism is identify the person accessing the card is an authorized one. This validation relies on the interaction with the ID card holder. The signature is the most common behavior used for identification. The signature "characteristics" of an identity is unique and unchangeable, so it is viewed as biometrics characteristics belonging to that unique identity. Biometric technologies are used to identify and authenticate the identity of personal unique physiological or behavioral characteristics. Since biometrics characteristics cannot be stolen or forgotten, they can provide a secure and convenient way of authentication for an individual. Other biometric technologies such as fingerprints, hand geometry, retina, Iris, and voice are possible. The biometric information can be stored in the ID card because chip-based ID card has much larger amount of memory. Adding biometrics information to Smart Cards improves personal security levels [9].

In short, the advantages of Chip-based cards include: 1) A Smart Card has powerful calculation, complicate logic control, and large storage abilities but easy to carry; 2) A Smart Card has several security mechanisms to protect inside data; 3) A Smart Card does not depend on a potentially vulnerable external resource; 4) A Smart Card's applicability allows multiple functionality; 5) A Smart Card has the

characteristic of cost-saving, particular for off-line application.

These positive features make chip-based national ID cards a favorite choice for governments when they try to issue new national ID cards or replace the current national ID cards. The security features of Smart Cards ensure the national ID card to prevent from fraud activities. The Smart Card has strong security, it is suitable for offline verification, it has strong information storage capacity, and it has multiple services and applications [10]. These security features can make the national ID card more reliable and more difficult to be forged. The chip-based ID cards can be good tools to implement secure personal identification systems. In short, the chip-based ID cards provide confidence in verification [11].

In addition to the U.S. government, two specific groups such as the military and Airline Company requiring higher security tend to favor the issue of chip-based ID cards [12]. In the aftermath of the 911 terrorist attacks, not only the Bush administration, but also the CEO of the Oracle started to appeal the general public to establish a National ID system so as to make the society more secure [13].

3. Controversial Issues about Chip-based National ID Card

A Chip-based ID card is never a single idea, and it involves a great number of issues. A centralized database could easily become a single point to attack, and the privilege of database use is difficult to determine. The civic privacy and anonymous liberties seem to be inevitably infringed with such a national ID system. Even the centralized system is secured from external attacks, it is difficult to completely prevent from various problems stemming from human variables, such as internal threats by management personnel.

In the technological aspect, issuing chip-based ID cards involves some other equipment such as card readers, PC terminals, database servers, as well as network connections, rather than the cards themselves. The frequent malfunction of these related equipment and the rapid renewal with the quickly changed technological environment contribute to major problems for the chip-based Smart ID cards. The feasibility of national ID cards to counter terrorism, crime, or illegal immigration is still doubtful. When it comes to the biometrics information that chip-based cards may contain, the skills of data collection or verification are not well developed yet; at the same time, adding biometrics data such as fingerprints to the chip-based ID cards does not necessarily ensure the card security.

The attempt to issue multiple purposed ID cards in certain countries to establish full-dimensioned e-governments brought new problems. Making all information about personal health, insurance, and finance completely transparent in a chip-based ID card not only infringes civic privacy, but also increases the possibility of fostering prejudice in the society. The next section of this paper is to discuss the controversial issues listed above in an attempt to clarify the cons of chip-based national ID cards.

3.1 Database Problem

To establish a national wide integrated ID card system inevitably requires unique identifiers such as numbers to present identities in the database system, and each citizen would be assigned a specific number by the government to present himself/herself. The “one unique number or identified per person” becomes a tool for the government to control or watch its people. Second, a centralized ID card database system could easily become a single point to attack because it contains nation-wide information. As long as the information in the database is valuable, the identity-related credentials are likely to be theft or misused by intruders such as terrorists, organized crime, or blackmailers. The database becomes a target for subversion [14].

Third, it seems difficult to grant different levels of access privileges. Distinct users should be able to access civic information in different levels in accordance with their needs. It is hard to determine which departments or agents have the privileges to read, create, or update the data existing in the database. Who should determine the privileges depends on who is the owner of the information. Whether the government has the right to decide the privilege for individuals if the information belongs to them is a major controversial point. Fourth, the Internet connection is essential to access the central database because the database must provide remote access when a piece of data needs to be verified. A duplicate database would be needed when an on-line verification is not feasible. At this time, the duplicate database becomes a vulnerable point to be attacked and controlled.

3.2 Civil Liberties: Invasion of Privacy and Anonymity

Although it seems easier for intelligence and law-enforcement agencies to share information, a centralized, chip-based nationwide identity system could also carry significant risks. The invasion of civic privacy and the freedom of anonymity become major

attacking points for those who doubt the feasibility of chip-based ID card systems.

In the eyes of opponents, the establishment of national ID system truly leads to the invasion of personal privacy. They argue that if the system were adopted, citizens would be under indirect watch, which is an invasion of a person's privacy. Privacy International, a human rights group as a watchdog on governmental and corporation surveillance, summarized four statements against the National ID Card system. "First of all, race, politics, and religion were often at the heart of older ID systems. Second, ID cards facilitate an increase in police powers. Third, ID cards' usefulness to police has been marginal, although Law and Order is a key motivation for the establishment of ID cards in numerous countries. Fourth, the privacy implications are profound." [15] Privacy International also noted that the chip-based national ID cards systems could put civic activities under unnecessary surveillance. Because of the huge capacity of such a card and the detailed records of every movement in the central system, the freedom of anonymity is seriously infringed.

3.3 Information Management and Insiders Threat

Once the database and privilege grant are established and technological security is not a concern anymore, human factors stemming from two aspects still affect the feasibility of chip-based ID card systems. The two factors include the unintentional mistakes by data administrators and intentional misbehavior by internal staff. It is impossible to prevent from any data entry mistakes, and database errors in the system are always possible. Unless a correction mechanism is established, the risk of inaccuracy would exist all the time. On the other hand, certain misbehavior by database users such as data disclosure also degrades the function of national ID systems.

3.4 Technological Challenges

A complex and comprehensive national ID system relies on multi-cooperation from personnel and technologies. The more entities components are involved, the more threats /risks might take place in the system. The chip-based ID card is not the only agenda in establishing the system. The card reader, PC terminal, databases, and network connections are so essential to the entire system, but the malfunction of these components could happen at any time. The reader might not be able to read the data from the card, the reader could destroy the data on the card, or the network connection is not stable or even down. To

keep them work well, regular maintenance is necessary, and it may cost more to upgrade because the lifetime of computer system is short. Currently, the most acceptable contact type Smart Card lifetime is around three years for normal frequently uses in a humid country. The lifetime can be substantial reduced if the weather becomes dry when the static current possibility increases. On the other hand, Smart Card itself currently is not entirely secure, and any technique to break the Smart Card is possible. Firewall protection, virus protection, Trojan house, and hacker invasion prevention are to be established, whereas other security measures to check the authenticity of IDs and verify the presenter are also necessary. Security audit, policy establishment, and policy enforcement are essential to deal with technological challenges.

3.5 Terrorism, Crime, or Illegal Immigration Prevention

In the aftermath of Sept. 11, secure ID card systems are considered as a tool to enhance security. The appeals to issue chip-based ID cards with a view to track terrorists and improve the security within the country are increasing, even though it involves the issue of privacy infringement. Issuing chip-based national ID cards is being considered as a quick approach to prevent from terrorism [16]. Once the approach is adopted, citizens, legal residents, aliens, and international students would be required to carry and present IDs, and it would be supposedly easier for the authority detect terrorists. However, the effectiveness of these national ID cards to stop terrorism, crime, or illegal immigration remains controversial. First, the countries with national ID cards still cannot prove that they can detect terrorism. National ID cards would allow the government to monitor the American populace more efficiently. Second, those terrorists could get ID cards by legal or illegal approaches and evade from police tracking. Unless the authorities have identified an individual as a threat, Smart Cards will not help much. Such an ID card system could track identified terrorists upon their interaction with certain agents and speed up the arrests, rather than lead to those unknown crime [17].

Third, the national ID system becomes an investigative tool to normal people only, but might not be able to find anything. Before the authority detect illegal crime, the ID system might have affected most law-abiding citizens and bring so much inconvenience to their daily life. People without IDs might be stopped boarding planes, writing checks, or going to the movies, but the tricky bad guys could do anything illegally with stolen IDs [18]. Fourth, the effectiveness of ID cards on illegal immigration is doubtful. A report by John J. Miller and Stephen Moore in 1995 concluded that the purpose of countering illegal immigration of the ID

system might be expanded to others and the government would need to spend a lot on system administration. The price for the ID system includes not only civic liberties but also huge financial costs, while the achievement on illegal immigration is still unsure [19].

3.6 Biometrics

The unchangeable feature makes biometrics a unique way for identification. It seems logically authentic and convenient to identify a person by signature, fingerprints, hand geometry, retina, iris, or voice. However, there are some controversial arguments concerning the application of biometrics on ID cards. First, biometrics is not a mature technology, and it still has problems to catch, store, read, or conduct identification. The technology is far from being proven, and many problems concerning automated schemes, system outages and database errors still need to be dealt with. Second, to impose the identification procedure on organs changes the relationship between human beings and their bodies. The sense of body possession and individuality might be dramatically changed because of the identification schemes [20]. Third, the adoption of biometric identification does not increase security. Obtaining an ID card requires only low-security documents such as a convincing passport or birth certificate, and it seems not difficult at all for those who attempting to get an ID to achieve their goals illegally. After all, biometrics only helps verify the cardholders from the card data, rather than examines clearly the authenticity of the card [21].

3.7 Problems with Multi-purposed Card

Some countries issue multi-purposed ID cards aiming to promote digital signature, IT activities, and e-commerce/banking/purse activities. These multi-purposed cards store not only personal information but also medical records, financial information, etc. The concept is very aggressive and ambitious, but it is extremely difficult to implement. First of all, the more functions or personal information are on the card, the more transparent one could be. Second, the ID cards with multi-information could foster the formation of discrimination in the society. Not only one's race, origin, and nationality are shown on the card, but also his/her financial, medical, and other data are implicit. How could we ensure that no embarrassment or prejudice could happen when officials or card-checking people such as job market employees deal with the cardholders? Three, the multi-purposed ID card also has access control problem. Granting access privileges for different types of data administrators is

never an easy task. The mechanism with different categories and files involves various government and private agencies. The more sectors are involved, the more difficult it would be to determine the access privilege or establish a protection mechanism. Last but not the least, storing both static and dynamic data in the multi-purposed cards could lead to interferences among different types of data or even complete malfunction of the card. Static data such as personal identification data is rarely updated, whereas dynamic data such as medical records and financial information is frequently changed. Once there is something wrong with one application, the other parts would be affected because the card itself could be blocked. The multi-purposed card does not function as well as expected, but causes more inconvenience.

4. Case Study: Chip-based National ID System in Asian Countries

In this part, Asian countries attempting to issue chip-based national ID systems are discussed. Most of these countries have already had national ID systems before talking about the chip-based ones. Establishing an E-Government, promoting e-business, and developing high technology are top priorities for these authorities. Development and prosperity of the whole country seems much more essential at this time than the appeals to think highly of individuals. For the administration convenience, certain countries even consider about issuing multi-purposed cards to contain all types of individual data in one smart card. However, the question that whether it is necessary to sacrifice individual privacy and interest to achieve the prosperity that is still blurred and indeterminate keeps challenging these governments in the process of consideration and adoption of such a chip-based system. Some governments cannot resist widespread question and opposition from the society and then decide to give up, while others insist executing the policy with elegant and grand reasons or purposes. In the next paragraphs, the process and results of practicing the chip-based ID card systems in Asian countries are briefly summarized.

Some countries including Singapore, Taiwan, Hong Kong, China, Malaysia, and Thailand have had a national ID system for years, and others such as Japan and South Korea have a residential registration system. Taiwan and China even have both systems. The residential registration system is used to keep track of people's movement and household status, although opponents argue that it forms another type of government surveillance in the society. In Taiwan and Malaysia, people need to carry their IDs all the time. When it comes to the new smart card technology, some

governments are still working on adopting the chip-based system, and widespread opposition and public pressure have stopped others.

Singapore has the system of digitalized bar code cards for year, and its citizens are required to have their ID cards. The government and military have fully deployed the chip-based ID card system, and the policy is still in progress. Hong Kong's national ID system was established in 1982, and everyone has been required to have his ID card. In November 1999, the authority established a multi-purposed Smart ID System. Citizens' residential status needs to be recorded on the card to prevent illegal immigration from China. The small population of 6.8 million makes the cost pretty low so that the policy can continue. China's 2nd generation ID card program has 840 million users and claims to be the world's largest smart card program [22]. It has multi-purposes, functioning with fingerprint, digital certificate, and other biometric features. The contactless smart card seems quite convenient for official investigations, and the authority had little privacy concern in adopting the system. Several cities such as Beijing and Shanghai have already used the tool to control floating population. That the authority faces rare public opposition and question in the society is the main reason that the policy continues. Malaysia's ID system was established in 1960, and all citizens have been required to carry the paper-based ID cards all the time. The authority even proposed a multi-purposed card named "Multimedia Super Corridor" as a tool to lead the

national development and plans to adopt the system overwhelmingly in 2003. Philippine is a particular case. The Supreme Court turned down the national ID system proposal in 1997, but the idea to issue chip-based ID cards arose again in May 2002. The authority is planning to establish a multi-purposed smart card system and add fingerprints to the card. The country's Senate is even working on legislation.

Contrast to the above countries, fate of the chip-based national ID card system seems unfortunate in Taiwan and South Korea. Taiwan's "All-in-one Citizen Card" proposal appeared in June 1998, and the proposal has been accompanied by the policy of "Build, Operate, and Transfer." Private companies promoted the proposal, and no capital allocation came from the government. The proposal failed in the end because of the broad critics and question in the society about the tendency of business-focusing policy and the potential risks of data leakage. Only a county government in an off-land island issued chip-based national insurance and health cards, despite the quite unsatisfying feedback to the new system from the county members. South Korea had its residential registration system established in 1968, and each citizen has been required to carry his residential registration card. The authority planned to issue a multi-purposed "Electronic National Identification Card" with fingerprints on it, but the plan was stalled and dropped in 1998 resulting from widespread concern about privacy infringement and human rights violation.

Table 1: Chip-based National ID card in Asian countries

Country	National ID Scheme	Compulsory	National ID Card	Chip-based ID Card	Result
Singapore	Yes, ID	Compulsory	Yes, digitalized bar code cards for year	Yes, government and military fully deployment [23]	Continuing
Taiwan [24]	Yes, ID/residential system	Compulsory	Yes, paper-based ID card, citizen need to carry it all the time	Yes, "All-in-one Citizen Card" in June 1998, "Build, Operate and Transfer" Government doesn't need money	Fail, "BOT" focus only business, only chip-based NIH card, survey said very unsatisfied to new system
Hong Kong [25]	Yes, ID (established in 1982)	Compulsory	Yes	Yes, "Smart ID System" in November 1999, Multi-purpose, residential status	Continuing, only 6.8 million population, low-cost, to prevent illegal immigration from China
China [26]	Yes, ID/residential system	Compulsory	Yes	Yes, "2 nd generation ID card", world's largest smart card program [27], 840 million users, multi-purpose, fingerprint, digital certificate, contactless smart card,	Continuing, no privacy concern needed, several cities such as Beijing, and Shanghai already using, tool to control floating population [28]

Thailand	Yes, ID	Compulsory	Yes	Yes	N/A
Japan	Yes, residential system	Compulsory	Optional	Yes, Optional	N/A
Philippine	No, Supreme Court shot down ID system proposal in 1997	N/A	Yes	Yes, even the Supreme Court decision, but arise again in May 2002, Multi-purpose, fingerprint	Continuing, country's Senate is working on legislation
Malaysia [29]	Yes, ID system (from 1960)	Compulsory	Yes, paper-based ID card, citizen need to carry it all the time	Yes, Multimedia Super Corridor, Multi-purpose card	Continuing, national roll-out in 2003, a tool to new leading edge homegrown technologies
South Korea [30]	Yes, residential system (from 1968)	Compulsory	Yes, residential registration card	Yes, "Electronic National Identification Card", Multi-purpose, fingerprint	Fail, stalled and dropped in 1998

5. Suggestions and Potential Solutions

The advantages of Smart Cards listed in this paper such as powerful calculation, strong storage abilities, and multi-functionality make the chip-based cards a great choice for governments to establish a national ID system. While the legitimacy and feasibility of a central ID system is still under intense debate, the fate of chip-based ID cards that store much more data and seem to infringe more privacy is not for sure. Various problems concerning the database, the invasion of privacy, information management, technological challenges, and biometrics make the society anxious about adopting the new system. The effectiveness of the chip-based ID system to counter terrorism and unlawful activities is doubtful, while so many new difficulties with multi-purposed cards take place in certain countries. To relieve the anxiety and concern about chip-based ID cards in the society, any government attempting to improve efficiency and effectiveness by establishing such a system should offer a set of comprehensive and convincing solutions.

The government should first make its policy, laws, and operation transparent to the public. The huge economic cost and the challenges of design and implementation make public review so important. The government must deliberate people's thoughts and clearly identify the goals of such a system. Privacy infringement and data leakage by hacking are people's major concerns about such an ID card system. Four solutions might be worthy to be considered in this aspect. First, the government might need to consider about storing personal data in personal cards only, rather than storing in a centralized database. Second, the data inside chips should be transparent to card holders. The right of self-information theoretically

ensures cardholders' privilege to check all the data inside the chip.

Third, a single card containing different types of personal data is never good for civic privacy. If the government is trying to establish nation-wide E-Systems to integrate national files, at least four types of cards such as personal ID, medical data, financial data, and security check should be issued separately to prevent from privacy infringement, discrimination, and interference among different functionality. Fourth, the authenticity of the chips is as important as that of the reading devices. Not only the reading devices should verify personal chips, but also could individuals verify the reading devices. Cardholders could trust the system only after confirming that their information would not be leaked or stolen by others [31].

To avoid from potential misidentification or misuses of the chip-based ID system, the government should design a strict scrutiny and supervision system. Regulations about the use control of such a system and punishment for misuse should also be enforced [32]. In addition, the government needs to make decisions about the particular agencies in charge of various activities in the ID system implementation. At the same time, an emergency security recovery system should also be established to deal with the problems when the system is failed to function or being compromised.

If the government aims to achieve long-term success of digital economy by establishing a chip-based ID card system, public values and public interests should be added into government consideration in advance of policy implementation. The huge economic costs and potential problems with the adoption of such a system will keep challenging a government all the time. If a government fails to gain public support or understanding prior to policy

implementation, it is inevitably for the government to be questioned on the policy motivation by its people. Even, the civics may not trust the government anymore, which is not a good sign for any government.

Smart Cards technologies may make it more difficult to forge ID cards, and the digitalized data on the cards may bring governmental effectiveness and efficiency and people's convenience. However, they are unlikely to provide a complete solution. The most

controversial issue of privacy by human rights groups is one of the multiple issues concerning the chip-based ID card system. In fact, it involves much more than privacy. Economic costs, security mechanism, administration, as well as legislation make the price too high to adopt such a system.

References

- [1] Privacy International, Aug 24 1996, "Identity Cards: Frequently Asked Questions", 2. What are the main purposes of ID cards?
http://www.privacy.org/pi/activities/idcard/idcard_faq.html#2
- [2] Bao Vu, "National ID",
<http://anegada.cudenver.edu/downloads/baovu/BaoVu.htm>
- [3] 'Identity cards- Frequently Asked Questions', Privacy International,
http://www.privacy.org/pi/activities/idcard/idcard_faq.html
- [4] Computer Science and Telecommunications Board (CSTB), 2000, "Ids- Not That Easy: Questions About Nationwide Identity Systems", chap 3, page 36-37,
http://www.nap.edu/html/id_questions/ch3.html
- [5] 'Identity cards- Frequently Asked Questions', Privacy International,
http://www.privacy.org/pi/activities/idcard/idcard_faq.html
- [6] Oliver Sylvester, "Types of Credit Card Fraud",
http://www.ex.ac.uk/politics/pol_data/undergrad/owsylves/page3.html
- [7] 'The View of The Asia Pacific Smart Card Association on the Proposed Hong Kong Special Administrative Identity Card'. The Asia Pacific Smart Card Association.
<http://www.legco.gov.hk/yr00-01/english/panels/se/papers/b204e02.pdf>
- [8] 'Understanding and Specifying Comprehensive Security for High-value/High-risk Smart Card-based Systems' Schlumberger White Paper.
<http://www.smartcards.net/pdf/sishell.pdf>
- [9] 'Smart Cards and Biometrics in Privacy-Sensitive Secure Personal Identification Systems', Smart Card Alliance. (May 2002),
http://www.smartcardalliance.org/pdf/alliance_activities/Smart_Card_Biometric_paper.pdf
- [10] 'Secure Personal Identification Systems', Smart Card Alliance. (January 2002),
http://www.smartcards.net/pdf/secure_id_white_paper.pdf
- [11] 'Smart Cards and Biometrics in Privacy-Sensitive Secure Personal Identification Systems', Smart Card Alliance. (May 2002),
http://www.smartcardalliance.org/pdf/alliance_activities/Smart_Card_Biometric_paper.pdf
- [12] Megan Lisagor (May 27, 2002), 'FAA preps Smart Card contract', Federal Computer Week.
<http://www.fcw.com/fcw/articles/2002/0527/news-faa-05-27-02.asp>
Christopher J. Dorobek and Dan Caterinicchia, 'CAC Identity Crisis', Federal Computer Week.
<http://www.fcw.com/fcw/articles/2002/0916/intercepts-09-16-02.asp>
- [13] Larry Ellison's 'Digital ID can help prevent Terrorism', October 8, 2001.
<http://www.oracle.com/corporate/digitalid.html>
- [14] Computer Science and Telecommunications Board (CSTB), 2000, "Ids- Not That Easy: Questions About Nationwide Identity Systems", page 34,
http://www.nap.edu/html/id_questions/ch3.html
- [15] 'Identity cards- Frequently Asked Questions', Privacy International,
http://www.privacy.org/pi/activities/idcard/idcard_faq.html
- [16] Pramod Shrestha, March 06, 2002, "National ID: How effective?"
<http://anegada.cudenver.edu/downloads/Pramod/nationalid.html>
- [17] Computer Science and Telecommunications Board (CSTB), 2000, "Ids- Not That Easy: Questions About Nationwide Identity Systems", page 5-6
- [18] October 13, 2001, "National ID card won't stop terrorists, but will infringe on Americans' liberty",
<http://www.lp.org/press/archive.php?function=view&record=539>
- [19] John J. Miller and Stephen Moore, September 7, 1995, "A National ID System: Big Brother's Solution to Illegal Immigration", <http://www.cato.org/pubs/pas/pa237.html>
- [20] Simon G. Davies, 1994, "TOUCHING BIG BROTHER: How biometric technology will fuse flesh and machine",
<http://www.privacy.org/pi/reports/biometric.html>
- [21] Computer Professionals for Social Responsibility (CPSR), Nov 27, 2001, "National Identification Schemes (NIDS) and the Fight against Terrorism: Frequently Asked Questions",
<http://www.cpsr.org/program/natlID/natlIDfaq.html#Q4>
- [22] Faulkner and Gray, Card Technology Cover Story, "China: The Smart Card Sleeping Dragon",
<http://www.advanceic.com/news.html#1>
- [23] Privacy International, August 24 1996, "IDENTITY CARDS, Frequently Asked Questions",
<http://www.totse.com/en/privacy/privacy/idfaq.html>
- [24] Ching-Yi Liu, 1999, "How Smart Is the IC Card? : The Proposed National Smart Card Plan, BOO Strategy, Electronic Commerce, and the Emerging Danger to

-
- Online Privacy in Asia”,
<http://www.isoc.org/inet99/proceedings/posters/439/>
- [25] D. Wiggins, Gartner Research, 15 February 2002,”
Hong Kong’s Multiapplication Smart ID Card”
LM Cheng, Smart Card Design Centre, “Comments on
Hong Kong Special Administration Region Identity Card”
- [26] PACE Integration, "Build it and they will come",
<http://www.paceintegration.com/articlesDetails.cfm?infobaseID=13>
Louisa Liu, Gartner FirstTake, “China’s Smart-Card IDs:
Economic Benefits, Increased Enterprise Risk”
- [27] Faulkner and Gray, Card Technology Cover Story,
"China: The Smart Card Sleeping Dragon",
<http://www.advanceic.com/news.html#1>
- [28] SMART CARD MARKETING, March 2002, “CHINA
SMARTCARD NEWS”,
<http://www.fondcard.com/magazine/news/>
- [29] Adam Creed, Newsbytes, Jul 2000, “Malaysian
Government Smart Card Project Gets Under Way”,
<http://www.fis.utoronto.ca/research/iprp/sc/malaysia06062000.html>
AP, THE AGE, March 11 2002, “Malaysia digital ID
slow to take off”,
<http://www.theage.com.au/articles/2002/03/11/1015365760982.html>
- [30] Choi Byoung-hw, “Resident Registration Card,
What's Up with It?”
<http://hjournal.hanyang.ac.kr/267/eureca.htm>
- [31] Roger Clarke, ‘Chip-Based ID: Promise and Peril’.
(September 1997),
<http://www.anu.edu.au/people/Roger.Clarke/DV/IDCards97.html>
- [32] Larry Ponemon, ‘Deciphering the security ID debate’.
(February 11, 2002), <http://news.com.com/2010-1078-833684.html>



Mobile Technologies, providing new possibilities in Customer Relationship Management

D.Arunatileka, MBA, B.Sc.
School of Computing & IT, University of Western Sydney
Locked Bag 1797, Penrith South DC
NSW 1797, Australia

darunati@cit.uws.edu.au

B. Unhelkar, Ph.D., FACS
School of Computing & IT, University of Western Sydney.
Bhuvan@cit.uws.edu.au

Abstract

CRMS is fast becoming a de facto standard for customer oriented technology and service organizations. The e-CRM creates a shift, providing customer service by ensuring a unified front of products and services to the customer. Mobile CRM, with real time customer interactions, pushes the CRM concept to its highest plane. The one-to-one personalisation has advanced into many to many relationships, bringing out the concept of i to i (individual to individual) which is achieved by product and services made available to a large group of mobile users at an unpremeditated time and spot. This paper addresses the possibilities offered by mobile technologies that promise to take CRM systems in to a new paradigm.

Keywords: CRM, CRMS, Mobile Technologies, e-CRMS, m-CRMS

1. Introduction

Siegel (1999), in *Futurise your Enterprise*, makes an emphatic point: "Over the next ten years, the Internet will drive changes in consumer behaviour that will lay waste to all the corporate re-engineering and cost reduction programs that have kept so many MBAs and programmers burning the midnight oil." While Customer Relationship Management Systems (CRMS) are a response to the changing demands placed on businesses due to the ubiquitousness of the Internet, we believe that this landscape has already undergone another massive change. This is because the Internet is itself undergoing another tumultuous change; instead of being available through a physical wire, the connectivity afforded by today's Internet is totally independent of

place: it is *wireless*. Personalisation has been the key to the rapid rise of CRMSes, as, based on the internet infrastructure, they have the abilities to provide personalised services to customers independent of time and place. This is because most CRMSs are internet-enabled, and are able to provide information, transaction and related facilities to both individual and business customers. Mobile technologies are shifting that paradigm to a new plane.

According to (AMR Research 2002), the market for CRM software was only US\$ 7 billion in 2000 which is predicted to increase more than threefold: up to US\$ 23 billion in 2005. With the advent of mobile technologies, these same CRMS functionalities are now being made available on the mobile screens of the customers, in trains, buses, shopping malls, schools, hospitals and almost any place that has mobile coverage. This has led to a revolution in terms of availability of services to customers. For example, SMS (Short Messaging Services) flight calls being made to select passengers provides enhanced service and time savings that would not have been possible by the most sophisticated CRM system that was connected by a physical wire. Other major areas of applications for mobile CRMS include airlines, hospitals, transport and retail. Mobile technologies have changed the landscape of CRMS applications. In this paper, we discuss in detail the differences between existing CRMS applications and the ones that are mobile-enabled. Based on these comparisons, a few valuable insights are offered in the way, new personalised customer-based processes can be designed and deployed on a m-enabled CRMS systems.

2. CRMS Landscape

2.1 What is CRMS?

CRMS, as the name suggests, is an attempt to integrate enterprise-wide technologies such as the front-end web site, the backend data warehouse, intranet/ extranet, call centre support systems, as well as the routine accounting, sales, marketing and production systems to provide a *unified* front of the business to the Customer. Thus, a CRM represents a strategically significant initiative by a business to convert the data and information it has on itself and on the customer, in to a personalised and valuable *relationship* with the customer. (Brohman et al., 2003) suggests that a CRM involves active use of and learning from the information collected from the customer.

(Madeja N. and Schoder, 2003) describe CRM as a revolving process during which companies interact with their customers, thereby generating, aggregating, and analysing customer data, and employing the results for service and marketing activities. Describing the concept further, (Madeja N. and Schoder, 2003) explain that the motivation for companies to manage their customer relationships is to increase profitability from concentrating on the economically valuable customers, increasing revenue (“share of wallet”) from them, while possibly “demarketing” and discontinuing the business relationship with customers of less business value. However this might not be fully achievable especially for service organisations, due to various other parameters such as regulations and legislation where certain groups of customers need to be served irrespective of their business value. For example, telecommunication service providers are bound by regulation to provide services to domestic customers in rural and remote areas, although the income generation from such sectors may not be economically so viable.

According to (Ocker and Susan, 2002) when considering a CRM initiative, executives ultimately want to know the impact of such an initiative on the organizational performance: that is, the likely business value of the initiative. This is typically measured by the return on Investment (ROI) metric. However, determining the economic value of an innovation, especially one enabled by technology, has posed major difficulties for several decades. Thus it is obvious that the actual measurement of the outcomes of CRM will take time to evaluate. However as the competition is vastly aggravated due to free market economies, globalisation due to Internet explosion, innovative marketing and business models, most organisations cannot wait to take the initiatives towards CRM based solutions.

Having argued for the urgency of a CRMS, let us briefly consider what constitutes the CRMS application. Typically, a CRMS is characterised by three significant

functions related to customers: Acquisition, Retention and Growth.

2.2 Acquisition

Customers acquisition is where new customers are acquired using the knowledge created through CRMS. This could be in the way of presenting a better value to the potential customers in giving real time information, more focused customer approach, customisation, etc. As this is perceived as a higher value addition by the customer, it will provide the right attraction for the customer to be acquired by the business. CRMS can help in demographic analysis enabling networking amongst potential customers and help the business in attracting more customers. It should be noted that acquiring new customers is one of the costliest exercises in business world but an essential part in a growing organisation.

2.3 Retention

In today’s Corporate world, it is very crucial to keep the existing customers as the cost of migrating to other suppliers is only a click away and of minimal cost to the customer. For example, the competitive telecommunication service providers would provide various value additions on top of basic services in order to attract and retain the right profile of customers to their networks. These would include bundling of services with percentage discounts for multiple service provision, Cross subsidising various services in order to offer primary services, adding different features as well as creating various permutations and combinations of existing products to suit the needs of the customers at minimal additional cost and different loyalty programmes. Moreover, this is very important as research suggests that a 5% increase in retention could mean a profit of many more times since retention is much less expensive than acquisition and a good customer retained may make twice more revenue than an unknown new customer.

2.4 Growth

Customer enhancement is adding value to existing customers. For instance, a customer having one service from an organisation would be given an overall discount on the event if the customer wants to activate a second service from the same organization. For example a telecommunication service provider who is also an Internet service provider could offer a discounted overall bill for a customer who has a telephone service, to get the customer to subscribe for Internet as well. The customer gets the benefit of getting an overall discounted bill and also a combined bill which is beneficial and convenient to him. Further, if the

customer is interested in mobile telephony, a further concession could be granted. The CRMS would handle such customers in a different way so that at every bill run, the customer gets a discounted combined bill. This also provides another advantage to the business organisation since it has the knowledge to choose the right customers for enhancement.

All in all, CRM is a strategic tool to acquire, retain and enhance customers. Increased competition due to free market principles and intensified customer demand fuelled by more knowledgeable customers also adds more pressure to offer better customer service.

3. E-CRM

3.1 Relating CRM to e-Business

The emergence of e-business gives rise to a new dimension in CRM that can be known as e-CRM. This concept facilitates the capture, integration and distribution of data gained at the Organization's web site throughout the enterprise (Pan and Lee, 2003). Metagroup predicts that the e-CRM craze will only intensify, with the market growing from US\$ 20.4 billion in 2002 to US\$ 46 billion by 2003 and perhaps to US\$ 125 billion in 2004. On the downside, Gartner group reports indicated that more than half of all e-CRM projects are not expected to produce a measurable ROI. (Fjermestad and Romano, 2003) states that the goal of e-CRM systems is to improve customer service, and to aid in providing analytical capabilities. Furthermore it is the infrastructure that enables the delineation of and increases in customer value, and the correct means by which to motivate valuable customers to remain loyal.

The traditional CRM has limitations in supporting outside multi channel customer interactions that combine telephone, the internet, email, fax, chat and so on. In contrast the e-CRM solution (front office suites) supports marketing, sales and service. Integration between CRM systems and Enterprise Resource Planning (ERP) systems is becoming more common. As also stated by (Fjermestad and Romano, 2003), e-CRM is a combination of hardware, software, processes, applications and management commitment. They further suggests that there are two main types of e-CRM: Operational e-CRM and analytical e-CRM. Operational e-CRM relates to customer touch points which could be inbound contacts through a telephone call or a letter to company's customer service centre or outbound contacts such as sales person selling to a customer or an e-mail promotion. Analytical e-CRM, on the other hand, requires technology to process large amounts of customer data. The intent is to understand via analysis, customer demographics, purchasing patterns, and other factors so as to build new business

opportunities. Thus it becomes obvious that e-CRM tends to be used in different ways, depending on the objectives of the organisation. E-CRM is thus a combination of technology, software and realignment of business processes with customer strategies.

3.2 Comparing CRM with e-CRM

Table 1 compares and contrasts the differences between CRM and e-CRM critically. The main differences occur in the Customer Service area where CRM offers target marketing for a particular group of customers, e-CRM could be very specific providing 1 to 1 marketing. The static, one way service within time and space limits are replaced by real time two way service which could be offered at any time, from any where.

	Customer Data	Analysis of Customer Characteristics	Customer Service
CRM	Data Warehouse Cust Information Transaction History Products Info	Transaction Analysis Customer Profile Past Transactions History	Target Marketing Static Service One-way service Time & Space limits
e-CRM	Web House Customer Info. Transaction History Products Info. Click Stream Contents Info	Transaction Analysis Cust. Profile Past Trans. History Activity Analysis Exploratory activities (Navigation, shopping cart, shopping pattern, etc.)	1 to 1 Marketing Real time service Two way service At any time From anywhere

Table 1 – Difference between CRM & e-CRM (Source : (Pan and Lee, 2003))

4. Mobile Technologies

Mobile technologies form the basis of the ‘next wave’ of software applications. Riding on the back of traditional internet, mobile networks ensure that information is available to its users independent of a physical location. This is a major wave after the internet, as users have a high expectation of mobile applications – particularly being able to access their applications from anywhere and at any time. Mobile technologies and the resulting mobile applications have brought about unique phenomena in personal, social and business domain that is unparalleled.

Some of the characteristics of the rapidly evolving new generations of Mobile technologies include:

- Higher speed access than earlier mobile devices – it is now possible to not only make phone calls but also browse the internet
- Integrated devices (e.g. Nokia 2000 communicator) consists of web browser, personal messaging, data organizing system and a phone – as compared with merely a phone and, say, a pager
- Wireless personal network – Bluetooth deliver network access anywhere for handheld gadgets, thereby connecting laptops to phones to desk machines and so on. A user can have a personal mobile network of her own in the house or at work.

4.1 Common Advantages of Mobile Technologies

Personalisation is the key to provide enhanced customer services in modern age. Mobile technologies facilitate that much better than any other technology. While Mobile technology is usually touted to increase costs that is not always true. For example, use of mobile technologies also results in improved efficiency through reduced cycle time and thereby reducing the cost effectively. A classic example is that of roadside services (such as one offered by NRMA) wherein the broken down vehicles notify the service provider of their location by mobile; and the system also supports the vehicle of the service provider crew by scheduling and directing them in the right sequence to the broken down vehicles. Logging of services and their completion can also be enabled through mobile gadgets. Similar methods could be used by taxi services since the mobiles multicasting would be more secure than radio phones. Mobiles could also cover a larger area and personalised information could be passed to the drivers if the customer was a regular.

4.2 Mobile Technology Statistics

Mobile devices by far has outnumbered the internet users and the personal computer users worldwide. There are twice as many mobile devices as compared to Personal Computer(PC)s. Therefore, one can deduce that the market opportunity for mobile applications is twice as much as that of PCs – at least for the leading provider of personalised services. Table 2 provides some global statistics on mobile/PCs and Internet users.

Service/product	1999	2000	2001	2002 (estimate)	2003 (Forecast)
Internet users(millions)	277	399	502	580	665
PC users(millions)	435	500	555	615	650
Mobile subscribers(millions)	490	740	955	1155	1329
Telephone lines(millions)	905	983	1053	1129	1210
International Telecom Traffic minutes (billions)	100	118	127	135	140

Table 2: Global statistics on telecommunications(Based: International Telecommunication Union (2001))

Based on the above statistics, it is obvious that the most popular personalised services in the future will be based on mobile devices that are currently growing at twice the speed of Personal computers and Internet. Thus a web which is mobile enabled would be the future for any CRM to make an impact on the customers.

(Varshney and Vetter, 2002) state that wireless and mobiles networks have experienced exponential growth in terms of capabilities of mobile devices, middleware development, standards and network implementation, and user acceptance. Currently, more than 800 million cell phones and other mobile devices are in use worldwide, and out of those, more than 140 million users are in US alone. As per the above table this figure has far exceeded. The worldwide numbers are projected to rise to one billion soon(which has been surpassed), thereby exceeding the combined total of all computing devices several fold. In addition, countries with a lack of regular telecom infrastructure are likely to adopt wireless and mobile communications to serve both urban and rural areas. Table 2 gives an insight into the statistics of Internet users, mobile subscribers and Personal computer users. Approximately 500 million of mobile users are predicted to be web enabled by the year 2003.

According to Gartner group, in 2004, at least 40% of business to consumer e-commerce will be initiated from smart phones supported by WAP(Wireless Application Protocol). A study from the wireless Data and Computing Service, a division of Strategy Analytics, reports that the mobile commerce market may rise to US\$ 200 Billion by 2004. The report also predicts that transactions via wireless devices will generate about US\$ 14 billion a year.

5. Moving towards m-CRM

5.1 m-CRM offerings

Many European CRM suppliers have already integrated support for WAP into their products. Some suppliers offer architectures designed specifically around mobile devices. Integrating wireless sales, marketing and support applications into contact centres is critical for organizations to optimize the customer experience and leverage all customer touch-points (Greenberg, 2001).

First of all, by extending sales applications to the user's customer location, enterprises can equip sales personnel with the necessary information to close a deal (PeopleSoft, 2002). Barratt (2002) also presents that customers today expect a rapid and personalized service in all interactions with a company, over any channel; and m-CRM enables sales, service and other frontline staff to meet these expectations.

It is also interesting to note how the mobile dimension extends the e-CRM to further enhancing the 1 to 1 marketing to 1 to many and many to many. This means that a Business to Business (B2B) relationships in e-CRM could be further atomised in the m-CRM to Business to Customer (B2C) relationship. The customers would be the process owners of the second business organizations. These process owners could be various different divisional heads etc. whose mobiles are enabled to receive messages from the system. Similarly, the organisation 1 also could have various personalities who could be authorised to interact with the organisation 2. Therefore it could also be classified as many to many relationship. However all these relationships are based on the requirements of the customer organisation and to that limit it could be personalised. Most messages could be personalised for relevant info for those different individuals. Most information could be sent in the form of SMS messages.

Therefore with the event of mobile technology coming into existence the possibilities on the customer relationships are enhanced manifold.

	Customer Data	Analysis of Customer Characteristics	Customer Service
CRM	Data Warehouse Cust Information Transaction History Products Info	Transaction Analysis Customer Profile Past Transactions History	Target Marketing Static Service One-way service Time & Space limits
e-CRM	Web House Customer Info. Transaction History Products Info. Click Stream Contents Info	Transaction Analysis Cust. Profile Past Trans. History Activity Analysis Exploratory activities (Navigation,shopping cart, shopping pattern,etc.)	1to 1 Marketing Real time service Two way service At any time From anywhere
Mobile CRM	Web House/Mobile enabled process owner details Customer Info. Individual process owner Info. Transaction History Products Info. Click stream Contents Info	Transaction Analysis Cust. Profile Individual process owner profiles Past Trans. History Activity Analysis Exploratory activities(Navigation,shopping g cart, shopping pattern,etc.) Individual preferences of customers identified	Many to many to Marketing (Mobile enabled) with the new concept of I to I (individual to individual) between business organisations Real time service Many way service At any time From anywhere Through mass customisation

Table 3 : Emergence of m-CRM through e-CRM

Note that the 1to 1 marketing has been further enhanced with many to many relationships where customers while having a personal contact with the Account Manager or the Customer care person, could also have a fruitful relationship enabled by the CRM possibilities and mobile robustness with many contact points or process owners allowing the customers to build up trust with the organisation.

As proved by statistics, mobile commerce is exploding with such rapid speed, it is inevitable that this would lead to a paradigm shift in the way business is conducted. New methods and applications would be developed to take the full advantage of this mobile explosion which would lead to new business models.

5.2 Comparison and Extension of e-CRM to m-CRM

Table 3 gives a detailed description of how e-CRM could be extended with the dimension of mobility.

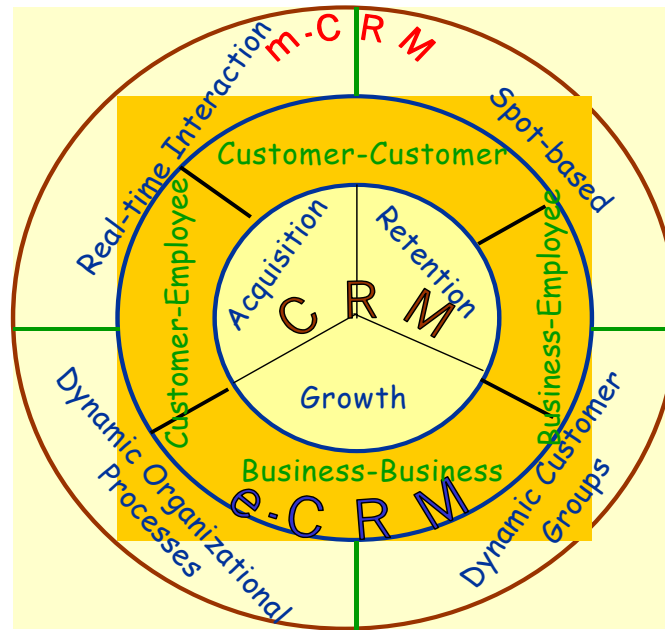
These applications would fit the limitations of current mobile devices but would enhance the business processes with real time customer service.

(Veijalainen et al., 2002) state, the main driving for the acceptance rate of small mobile devices is the capability to get services and run applications at any time and at any place, especially, while on the move. The experience from Japanese market shows that the most important factor is that the terminals are permanently carried around, and thus people can use so-called “niche-time” to use the gadgets for various things.

The number of these mobile Internet-enabled terminals, sometimes called Personal Trusted Devices (PTDs), is expected to exceed the number of fixed-line Internet users around 2003.

6. Original Value-add by m-CRM

Therefore, we take a view that m-CRM not only expands and enhances what is offered by e-CRM, which in turn expands and enhances on what is offered by CRM in general, but that m-CRM is more than merely making the CRM and e-CRM functionalities mobile.



This is being explained in Figure 1.

Figure 1: Original Value added by m-CRM over CRM and e-CRM

While acquisition, retention and growth have been traditional CRM functions, the e-CRM enhances these functions by providing ability for customers to interact with other customers. Excellence in customer service is achieved by bringing the relationship between the customer and the employee very close. However, the business also has the opportunity to improve its internal business processes by facilitating employee-employee relationship. And finally, Business to business communication and relationship is enhanced by e-CRM.

Mobile CRM further enables that which is provided by e-CRM. However, with mobility comes real-time interaction with the customer. Therefore, the four *additional* characteristics of m-CRM are:

- Real-time Interaction
- Spot-based
- Dynamic Organizational Processes
- Dynamic Customer Group

6.1 Real-time interaction.

As compared with traditional e-CRM, where the location of the customer is fixed, or is certainly temporarily from a physical connection, in case of m-CRM the interaction with the customer can be (is) in real-time. This is because the customer could be physically moving in real time and still, apropos certain legal and moral constraints, easily reachable. Real-time interaction offers customers to interact real time. The customer could be anywhere for that matter but could

interact through m-enabled CRM system. This would offer better value for customers since specific problems could be solved then and there without delays. The choice is at the customer's hand to attend real time or defer interaction for later. For example, a message (SMS) is received by a customer from a mobile-enabled parking machine indicating that the parking time is lapsing in so many minutes. The customer either could move his vehicle or re-charge his parking(using the mobile to dial into m-CRM).

6.2 Spot-based.

Wherein, the customer's physical location is of importance, and not just her virtual location. This provides unique opportunities to combine dynamically the physical presence of the customer with the significance of that customer in correlation with certain locations. A common example is a customer walking through a shopping arcade with mobile phone in her pocket, and she is provided with "spot sales" on hairdo and nail polishing – something that is difficult to sale on the land-based Internet. Another such instance is, mobile phone users within the periphery of a shopping centre receiving a SMS offering a 30 % discount at McDonalds within the premises on a particular burgher if the order is placed within the next half an hour.

6.3 Dynamic Organizational Processes

Change to organisational processes is when, mobility has forced the existing system to change into a new way of doing things. For example, the road services use the mobile technology to get the calls of customers needing road assistance and the vehicles on the road are notified by the system that a request has been made in a particular location. Then the road service provider notifies the system that he is within such a distance to the requested location and would take so many minutes to reach. Using this, Road service providers could be strategically located in locations where services are most needed so that the lead time to such services could be minimised. Also by doing this, efficiency is increased and the cost could be decreased, due to less travel by the service providers.

6.4 Dynamic Customer Group

Many new relationships is happening real time through messaging etc. irrespective of customer's location to a targeted group of customers.

Dynamic customer group creation and targeting is facilitated by Mobile technologies. Customer-group based services are where a selected group of customers are notified of a special service. For instance, the members of a frequent flier programme could be notified that a flight is delayed by so many hours due to some reason, so that anybody belonging to such club rushing for the flight could take it easier in reaching the airport. More importantly, a group of customers in a shopping mall change every day, but each day a new group can be dynamically created depending on the interests of the customers and, say, the excessive stocks in the mall.

7. Conclusion

Due to the predicted growth in the mobile market, applications for m-commerce are on the increase. The forces such as the high wireless installed base, emerging standards for wireless protocols, ever increase of bandwidth in the wireless networks also force the growth prospects of mobile commerce. CRM is becoming a de facto standard for organisations in order to compete effectively in today's global market. E-CRM which has extended the normal CRM systems to conduct business transactions electronically has improved the concept of CRM. M-CRM which synergises the dimension of mobility into e-CRM adds many new aspects to the e-CRM concept. Similar to the shift that happened when organisations had to undergo a paradigm shift in re-engineering the business process in order to support the e-economy, m-CRM also would make a significant shift in the business processes.

M-CRM offers four specific advantages in addition to those offered by e-CRM: real time interaction, spot based customer interactions, dynamic organisational processes and dynamic customer groups. Organisations would have to re-engineer all their support processes in order to implement a successful system to accommodate such new challenges. The need for such implementation would intensify with time due to competition. Each business organisation offering dynamic deals would be looked upon in a positive way by a knowledgeable market. The differentiation of service would offer the highest value in the eyes of the customers. Ironically, the organisations, wanting to adopt m-CRM as marketing and differentiation tool only, would have no choice but settle for re-engineered processes thereby change their business models in order to survive in the long run.

Acknowledgements

The authors wish to acknowledge the support of University of Western Sydney, School of Computing and IT, as well as Mobile Internet Research and Applications Group (MIRAG) within AeIMS research group at the school for their invaluable support in being able to present this paper.

References

1. Siegel, D., *Futurize Your Enterprise: Business Strategy in the Age of the e-Customer*, John Wiley and Sons, 1999; www.futurizenow.com.
2. Barratt, C. (2002), *Mobile Technology: Communications, Consultants' Advisory*, www.consultant.advisory.com/2002/june/cwticles.asp
3. Brohman, K. M., Watson, R. T., Piccoli, G. and Parasuraman, A. (2003), " Data Completeness: A key to effective Net-based Customer Service Systems" *Communications of the ACM*, **46**, 47-51.
4. Fjermestad, J. and Romano, N. C. (2003), " An integrative implementation framework for electronic customer relationship management: revisiting the general principles of usability and resistance" In *36th Annual International Conference on system sciences* IEEE, Hawaii, USA, pp. 183-191.
5. Greenberg, P. (2001), *CRM at the Speed of Light: Capturing and Keeping Customers in Internet Real Time*, McGraw-Hill
6. Madeja N. and Schoder, D. (2003), " Impact of electronic commerce customer relationship management on corporate success - results from an empirical investigation" In *36th Annual International Conference on system sciences*. IEEE, Hawaii, USA, pp. 181-190.
7. Ocker, R. J. and Susan, M. (2002), " Assessing the Readiness of Firms for CRM: A Literature Review and Research Model", In *36th International conference on System Sciences* IEEE, Hawaii, USA.

8. Pan, S. L. and Lee, J.-N. (2003)," Using a e-CRM for a unified view of the customer *Communications of the ACM*, **46**, 95-99.
9. PeopleSoft (2002), *The Business Case for Mobile CRM: Opportunities, Pitfalls, Solutions*, PeopleSoft White Paper Series,
http://www.peoplesoft.com/media/en/pdf/Mobile_CRM.pdf
10. Varshney, U. and Vetter, R. (2002) In *Mobile Networks and Applications*, pp. 185-198.
11. Veijalainen, J., Terziyan, V. and Tirri, H. (2002)"Transaction Management for M-commerce at a Mobile terminal", In *36th International Conference on System Sciences* IEEE, Hawaii, USA.



Optimized Web Information Retrieval with Fuzzy Logic

Harshana Liyanage
Dept. of Computer Science and Statistics
Faculty of Science
University of Peradeniya, Sri Lanka

G.E.M.D.C.Bandara *
Department of Production Engineering *
Faculty of Engineering
University of Peradeniya, Sri Lanka

Abstract

The World Wide Web contains a huge amount of unclassified data and its continuous growth has made it a complex domain for information retrieval. Current web information retrieval(IR) systems(i.e., Search engines) very often overload the user with irrelevant search results. This has forced the user to perform a certain level of analysis on the results returned. Web IR systems are currently one of the most researched areas in the computer industry. So far there have been many attempts to incorporate Soft Computing techniques such as Fuzzy Logic, Neural Networks, Genetic Algorithms, etc. This paper focuses on how Fuzzy Logic can be introduced to IR systems. The current applications of Fuzzy techniques are analyzed and a concept called “Macro-clustering” is introduced as a solution for optimizing results of generalized search queries.

Keywords: Data Mining, Fuzzy Systems, Fuzzy Clustering, Information retrieval, Fuzzy Searching, Proxy Design Pattern

1. Introduction

The Internet contains a vast collection of unclassified information. This information is heterogeneous in nature as it consists of text, images, graphics, animation, video clips, audio streams. The number of documents in the World Wide Web counts to several billion and the expansion of this scale has led the average web surfer tangled in volumes of unnecessary information. Such situations have forced researchers to build a more ‘intelligent web’. The study of Information Retrieval (IR) has become one of the driving forces in the path to this goal and most of the research conducted on this area has been on data mining the web. The following characteristics have made the web a complex domain for data mining operations.

1. Distributed data: Documents a spread over millions of different web servers

2. Volatile data: Many documents change or disappear rapidly (e.g. dead links)
3. Large volume: Billions of separate documents
4. Unstructured and redundant data: No uniform structure, HTML errors, Duplicated documents
5. Quality of data: No editorial control, presence of false information, typing errors
6. Heterogeneous data: Multiple media types(image, video, sound), languages, character sets

2. Problems in current web IR systems

There are many difficulties encountered during the retrieval of information on the web[1]. The key problems of current web IR systems are :

- *Imprecision, and Uncertainty*
- *Lack of deduction capabilities*
- *Inability to take soft decisions*
- *No page ranking with respect to user queries*
- *No personalization*
- *Dynamism, Scale, and Heterogeneity*

Imprecision, and Uncertainty: The aim of an IR system is to estimate the relevance of documents to users’ information needs, expressed by means of queries. This is a hard and complex task which most of the existing IR systems find difficult to handle due to the inherent imprecision and uncertainty related to user queries. Most of the existing IR systems offer a very simple modeling of retrieval, which privileges the efficiency at the expense of accuracy. Query processing in search engines, which are an important part of IR systems, is simple blind keyword matching. This does not take into account the context and relevance of queries with respect to documents

Lack of deduction capabilities: The current search engines have no deductive capability. For example, none of them gives a satisfactory response to a query like: How many computer science graduates were produced by South Asian universities in 2003?

Inability to take Soft Decisions: Current query processing techniques follow the principle of hard rejection while determining the relevance of a retrieved document with respect to a query. This is not correct

since relevance, itself, is a “gradual” property of the documents [2], not a crisp one.

Page Ranking with respect to user queries: Page ranks are important since human beings find it difficult to scan through the entire list of documents returned by the search engine in response to his/her query. Rather, one sifts through only the first few pages, say less than 20, to get the desired documents. Therefore, it is desirable, for convenience, to get the pages ranked with respect to “relevance” to user queries. However, there is no definite formula which truly reflects such relevance in top-ranked documents. The scheme for determining page ranks should incorporate 1) weights given to various parameters of the hit like location, proximity, and frequency; 2) weight given to reputation of a source, i.e., a link from yahoo.com should carry a much higher weight than a link from any other not so popular site; and 3) ranks relative to the user.

Personalization: It is necessary that IR systems tailor the retrieved document set as per users’ history or nature. Though some of the existing systems do so for a few limited problem domains, no definite general methodology is available.

Dynamism, Scale, and Heterogeneity: IR systems find difficulty in dealing with the problem of dynamism, scaling, and heterogeneity of web documents. Because of the time-varying nature of web data many of the documents returned by the search engines are outdated, irrelevant, and unavailable in the future, and, hence the user has to try his queries across different indexes several times before getting a satisfactory response. Regarding the scaling problem, Etzioni and Zamir[3] has studied the effect of data size on precision of the results obtained by the search engine. Current IR systems are not able to index all the documents present on the web and this leads to the problem of “low recall.” The heterogeneity nature of web documents demands a separate mining method for each type of data. (e.g. Image retrieval systems[4])

3. Fuzzification strategies in IR

The following section discusses various attempts made in order to incorporate a fuzzy logic to information retrieval techniques.

One of the earliest fuzzy based IR system, RUBRIC, was developed by R.Tong, V. Askman and J. Cunningham[5]. It was capable of retrieving based on the relevance of the document to the semantics of the users query. RUBRIC used rules to link words to concepts, which are also connected to semantically related concepts by rules. Each rule has a “relevance value” to indicate the strength of the association between its antecedent/ words and consequent

concepts. For example, if both “killing” and “politician” occur in a document, it suggests the document is somewhat related to assassination. This can be expressed as a rule with relevance value 0.5. During the information retrieval RUBRIC views the user’s query as a goal. Through goal back driven chaining, it determines the degree the document is relevant to the query. Fig 1. shows an example inference tree for finding documents related to the concept terrorism. The leaves that are found in a specific document are assigned a relevance value of 1. Otherwise, the term is given a relevance value of 0. These relevance values are propagated upwards in the inference tree using rules. Several rules have “auxiliary antecedents”, whose appearance in the document modifies (usually strengthens) the relevance value of the rule consequent. For example, a violent action is quite (0.8 degree) relevant to a terrorist event. However a violent action occurring together with assassination is completely (1.0) relevant to a terrorist event, as shown in the figure.

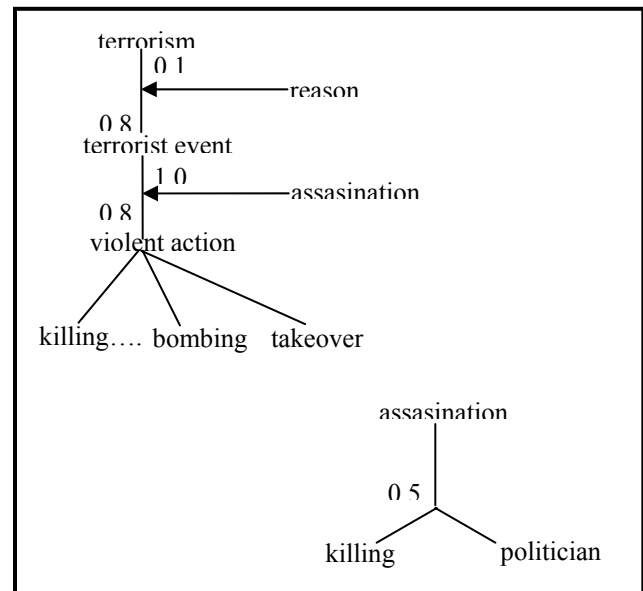


Fig. 1

The relevance of a inferred concept is calculated by RUBRIC based on the following equation:

$$v[\text{consequent}] = a \times v[\text{primary}] + (b-a) \times v[\text{auxiliary}]$$

where a and b are relevance values associated with the primary antecedent and the auxiliary antecedent respectively, $v[\text{consequent}]$ denotes the relevance of the current document to the concept c . Relevance values are concept inferred from multiple rules and combined using fuzzy disjunction.

Using a measure of *recall* (i.e., the ration of number of relevant documents in the database) and a

measure of *precision* (i.e., the ration of the number of relevant documents retrieved to the total number of documents retrieved). RUBRIC's performance in retrieving information has been shown to be superior to a comparable approach using nonfuzzy rules. Even though RUBRIC's approach is effective for retrieving documents related to a small set of predefined concepts, it is difficult to scale up the approach for general information retrieval systems. An alternative approach is to automatically generate a relevance measure between keywords by analyzing all documents in the database of an information retrieval system. S. Miyamoto proposed an approach for generating fuzzy relevance measure (called *fuzzy pseudothesaurus*) based on the assumption that if two terms occur together frequently in documents, then they are relevant[6]. These IR systems have become models for web based IR systems.

Retrieving desired information from the Web is tiresome process that every web user goes through every day. The main reason is the poor classification of the Web information. Different search engines use different techniques for this purpose and as a result their results significantly different. (e.g. the popular search engine 'Google' uses link analysis to obtain its result). Thus the web user has to use several search engines in order to fulfill a particular information need.

There are plenty of Web search engines which utilize special robots in search for new Web pages, and when a page is found it is put in the 'right' classification category depending on the classification method the Web search engine is using. In a technique called "Metadata classification" embeds the parameters required for classification in the web object itself. This means that the task of classification is partially transferred to those who create and maintain Web elements. As an advanced solution for Web classification M. Marchiori [7] proposed a fuzzification method. He says that existing Web metadata sets do have attributes assigned to objects, but they either have them or do not have them. Instead, he argues that attributes should be *fuzzified*, i.e. each attribute should be associated with a "*fuzzy measure* of its relevance for the Web object, namely a number ranging from 0 to 1". This means that if an attribute is assigned value 0 it is not pertinent to the correspondent Web object. If the value is 0.4 relevance of the attribute to the Web object is 40%. Since classification by itself is an approximation, better or worse, fuzzification method allows flexibility within a predefined classification system providing more detailed ranking and allowing the basic set of concepts to be relatively small.

Similarly, Yager has described a framework for formulating linguistic and hierarchical queries[8]. It describes an IR language which enables users to specify the interrelationships between desired

attributes of documents sought using linguistic quantifiers. Examples of linguistic quantifiers include "most," "at least," "about half." Let Q be a linguistic expression corresponding to a quantifier such as "most" then it is represented as a fuzzy subset Q over $I = [0,1]$ in which, for any proportion r , belonging to I , $Q(r)$ indicates the degree to which r satisfies the concept indicated by the quantifier Q . Koczky and Geddon [9] deal with the problem of automatic indexing and retrieval of documents where it cannot be guaranteed that the user queries include the actual words that occur in the documents that should be retrieved. Fuzzy tolerance and similarity relations are presented, and the notion of "hierarchical co-occurrence" is defined that allows the introduction of two or more hierarchical categories of words in the documents.

The usage of fuzzy logic on web IR has been demonstrated by Molinari and G. Pasi [10], when they developed a principled approach for assigning weights to different components, which are specified by tags in an HTML document. The rationale is that a word in the title carries much more weight than the same word appearing in other portions of the document. Therefore, it is possible to sort tags based on their degree of importance. For instance a sorted list maybe Title, Header1, Header2, emphasis, delimiters, etc. Based on the order in the list, fuzzy weight can be calculated for each tag.

$$w_i = \frac{(n - i + 1)}{\sum_{i=1, \dots, n} i}$$

where the n is the total number of tags in the sorted list. The total relevance measure of a document, denoted F , to a query q is a weighted sum of relevance measure for each tag.

$$F(q) = \sum_{i=1}^n W_{t_i} \times F_{t_i}(q)$$

where $F_{t_i}(q)$ denotes the degree that the content of tag t_i is relevant to the query q . The aggregation is just the summation of the $F_{tag_i} \times w_i$ for all the i values.

4. Back-propagation to utilize relevance

So far we have seen how to calculate relevance of a Web object to the search query providing the information that is contained within the Web object. Since Web is a dynamic media inter-wound with hyperlinks it is a common fact that one Web object points to some other Web object or even to more of them. This brings us to the problem of calculating relevance of an object that is pointed to by some other

object, as described in the following model by M. Marchiori [7].

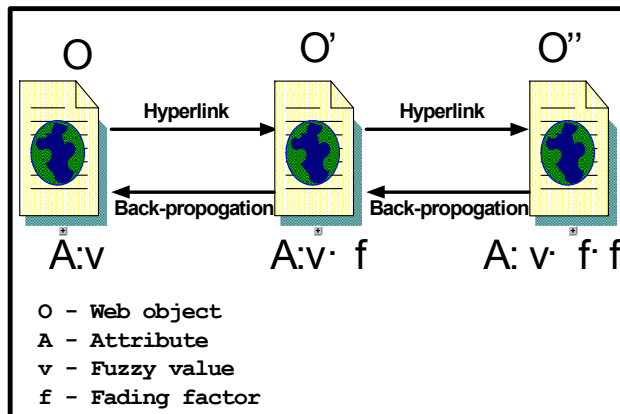


Fig. 2

Suppose a certain Web object O has the associated metadatum $A:v$, indicating that the attribute A has fuzzy value v . If there is another Web object O' with an hyperlink to O, then we can "back-propagate" this metadata information from O to O'. The intuition is that the information contained in O (classified as $A:v$) is also reachable from O', since we can just activate the hyperlink.

However, the situation is not like being already in O, since in order to reach the information we have to activate the hyperlink and then wait for O to be fetched.

So, the relevance of O' with respect to the attribute A is not the same as $O'(v)$, but is in a sense faded, since the information in O is only potentially reachable from O', but not directly contained therein. The solution to this problem is to fade the value v of the attribute multiplying it by a "fading factor" f (with $0 < f < 1$). So, in the above example O' could be classified as $A:v \cdot f$. The same reasoning is then applied recursively. So, if we have another Web object O'' with an hyperlink to O', we can back-propagate the obtained metadatum $A:v \cdot f$ exactly in the same way, obtaining that O'' has the corresponding metadatum $A:v \cdot f \cdot f$.

Experiments on a randomly chosen region of Web objects showed that the usage of back-propagation method can significantly improve effectiveness of the classification. They also showed that the critical mass of Web metadata classification usefulness is achieved when at least 16% of the Web use metadata classification, in contrast with 50% without incorporation of the back-propagation method. Furthermore, in order to achieve excellence level metadata need 53% of the Web to be classified, in contrast with 80% without described method.

Most of all, the method of back-propagation, which presuppose the fuzzification method, acts on top of any classification, and does not require any form of semantic analysis. Therefore, it is completely language independent which is very important when the number of non-English Web pages is constantly increasing.

5. Clustering of search results

Clustering techniques, in general, are used when there is no class of data to be predicted but rather when the instances of data are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instance to bear a stronger resemblance to one another than they do to remaining instances[16]. In the context of web information retrieval, clustering is used for automatically discovering groups of similar documents in a set of documents and a group of documents formed in the process is called a 'cluster'.

Clustering algorithms such as K-means, Buckshot, Fractionation, Suffix Tree Clustering are being used in existing web IR systems. Many of these document clustering algorithms rely on off-line clustering of the entire document collection, but the Web search engines' collections are too large and fluid to allow off-line clustering. Therefore clustering has to be applied to the much smaller set of documents returned in response to a query. Because the search engines service millions of queries per day, free of charge, the CPU cycles and memory dedicated to each individual query are severely curtailed. Clustering usually has to be performed on a separate machine, which receives search engine results as input, creates clusters and presents them to the user.

Clustering is important in several factors of information retrieval. In traditional information retrieval, one important means of speedup is to cluster data and to represent only a representative of each cluster in the database. [11]. When the source of information is the Internet, clustering the results allows more useful information to be presented on the first page of the results, allowing the user to determine which cluster is relevant.

The following search results illustrates the use of clustering in web IR. Here the keyword 'salsa' has been searched on the search engine WebCrawler[3].

Search results for the query: "Salsa"

Documents:246, Cluster:15

Cluster no.	Size	Shared phrases and sample document titles
1	8	Puerto Rico; Latin Music 1. <i>Salsa Music in Austin</i> 2. <i>LatinGate Home Page</i>
2	20	Follow Ups post; York Salsa Dancers 1. <i>Origin and Development of Salsa?</i> 2. <i>Re: New York Salsa Dancers are the best because...</i>
3	40	music; entertainment; latin; artists 1. <i>Latin Midi Files Exchange</i> 2. <i>Salsa Music On The Web. con Sabor!</i>
4	79	hot; food; chiles; sauces; condiments; companies 1. <i>Religious Experience Salsa</i> 2. <i>Arizona Southwestern Cuisine and Gifts</i>
5	41	pepper; onion; tomatoes 1. <i>Salsa Mixes</i> 2. <i>Salsa Q & A</i>
...

Table 1

Clustering was introduced to the web as a method of limiting the number of documents that a user is shown. An early experiment in Web document clustering has shown that allowing relevant documents to appear in multiple clusters is advantageous [12]. Fuzzy clustering [13] is a well known generalization of clustering where each element can have non-zero membership in multiple clusters. Cluster exemplars are then computed taking into consideration the relative membership of each member of the cluster. Given the complexity of the results of most internet searches, a fuzzy clustering is likely to better represent the data than a crisp clustering. In addition, the ability to represent and use the degree of membership in the cluster when determining cluster exemplars for display and relevance ranking will help to mediate the effect of cluster outliers that could prevent the user from seeing documents in a cluster that would otherwise be relevant. Google, which seems to be the most effective search engine to date, currently supports simple hostname-based clustering.

Etzioni[3] has listed the key requirements of web document clustering as measure of relevance, browsable summaries, ability to handle overlapping data, snippet tolerance, speed and incremental characteristics. In [14], fuzzy c-medoids (FCNdd) and fuzzy c Trimmed medoids (FCTMdd) are used for clustering of web documents and snippets (outliers). In [15], a fuzzy clustering technique for web log data mining is de-scribed. Here, an algorithm called competitive agglomeration of relational data (CARD) for clustering user sessions is described, which

considers the structure of the site and the URLs for computing the similarity between two user sessions. This approach requires the definition and computation of dissimilarity/similarity between all session pairs, forming a similarity or fuzzy relation matrix, prior to clustering. Since the data in a web session involves access method (GET/ POST), URL, transmission protocol (HTTP/FTP), etc., which are all nonnumeric, correlation between two user sessions and, hence, their clustering, is best handled using fuzzy set approach.

Although clustering improves the readability of search results, for general queries only a few large clusters are returned.

6. Macro-clustering

Web IR systems perform queries by considering web documents on an individual basis due to the architecture of the web. However, information is introduced to the web as websites having multiple related documents. This grouping formed by the creator of a web site is usually not taken into consideration by any IR retrieval technique. An objective of this paper is to demonstrate the use of a clustering method performed at the web site level.

As mentioned in the previous section existing clustering methods can improve the retrieval process by grouping results in a more meaningful format. On the other hand, for more generalized queries where the number of results in a cluster increases it is efficient to use a second level of clustering – known which the author names as “Macro-clustering”. This method takes in to consideration the possible classification of documents at a website level as shown in Fig. 3. A query submitted to a search engine can be of 2 forms.

A *General Query* is submitted by the user expecting a body of information. The number of occurrences of such information on the web is high and as a result a greater number of results are retrieved. On the other hand, *Specific queries* are submitted with the need for a specific piece of information. They produce in a fewer number of search results.

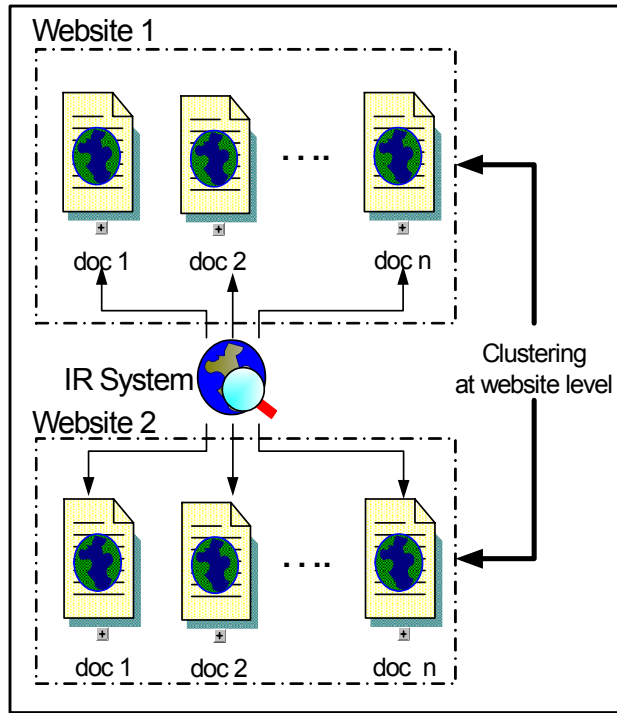


Fig. 3

Clusters formed in the process of a general query usually contains fewer clusters and a large amount of documents within a cluster. Such clusters tends to be less useful from the perspective of the user. For such clusters, by applying “Macro-clustering” we reduce the number of results in a number of 2nd level clusters (Macro-clusters) that are more manageable to the user.

The number of results retrieved by a query is proportional to the specificity of the query string. This can also be given as,

$$n_r \propto \frac{1}{n_k}$$

where n_r is the number of results and n_k is the number of key words in the query term. The most popular search engine to date, Google, accepts an upper bound of 10 key words.

For a more generalized query (e.g. $n_k < 4$) the impact of clustering is reduced and the number of results within a cluster increases. Also, for such queries (e.g. query string such as “Artificial Intelligence” or “Neural Networks”) it is difficult point out a specific document in a website that is related to the subject. For such situations, this second level of clustering will make the results of the query more useful. However, Macro-clustering may not be required for more specific queries (e.g. $n_k > 3$) as the results retrieved are more likely to be distinct documents in a website. As a result the decision to use this technique could be determined

by the IR system by analyzing the query string. Consider the following query strings.

$q1$ - “Fuzzy Logic”

$q2$ - “Applications of extension principle in fuzzy logic”

The above queries can be classified as generalized and specific query strings, respectively. By implementing a Macro-clustering the results retrieved for $q1$ should be mainly website based and, similarly document based for $q2$.

An experiment was carried out to distinguish the relationship between the number of query terms and the results retrieved. It was performed by executing search phrases $q1$ and $q2$, on several top ranking search engines. The results obtained are given in the following table.

Search Engine	No. of hits for $q1$	No. of hits for $q2$
1. Google	599,000	11,500
2. Alltheweb	294,095	649
3. Altavista	123,121	1,984
4. Wisenut	132,060	2,070

Table 2

These results are agreeable with the concept discussed above. Considering the given query phrases, a general query produces results in the order of 10^6 where as a specific query produces results in the order of 10^3 .

Relevance ranking can be incorporate into this method by using a fuzzification technique described in section 2 of this paper. By using a fuzzification technique we can assign a fuzzy value for the closeness of a document to a particular query phrase and subsequently using a clustering algorithm, document clusters can be formulated. The fuzzy relevance values of the documents can be used to index documents within a cluster.

With Macro-clustering we can perform a sorting of results within the cluster by calculating the aggregate value of fuzzy terms for each website. This can be found for the j^{th} website referenced in the cluster using,

$$F_{j \text{ aggregate}} = \sum_{i=1}^n F_{ij}$$

where $F_{j \text{ aggregate}}$ is the total of fuzzy values for a website and F_{ij} is the fuzzy value of the i^{th} document in the collection of n documents in the j^{th} website. In each cluster the website having the most number of relevant documents will appear at the beginning of the list.

The results of Macro-clustering will be a list of websites and upon selecting a website in the list, the user will be shown all of documents that match the query that belongs to the website. Here, the user will also be given the chance of visiting the home page of a website from where he could browse using the navigation system available in the site itself.

A method of classifying a website is required in implementing this method. The best method to achieve this is by performing link analysis within the search domain. Here, we can extract the domain information to identify documents coming from the same website. This information can be used as the basis for Macro-clustering. Here the user classification at the website level is deduced by the system by analyzing the URLs. An assumption is made here, that all web documents in a website have similar URL patterns. Another other method that could be used for identifying a documents parent website is to use identification that can be specified by the web document creator as meta data. For example, we can use the following notation to label documents in a website.

<META NAME= "website" CONTENT="www.fuzzylogic.com">

However, this method leaves a certain level of responsibility on the hands of the web document creator, and therefore, not very effective.

In the implementation Macro-clustering search engines would have to first determine the type of search performed based on the number of keywords. The following pseudocode shows a high level algorithm that can be used to process a query.

```

Accept user query phrase
Determine fuzzy values of documents in search domain
Determine query category
If 'specific-query'
    Cluster documents based on Concepts
Else If 'general-query'
    Cluster documents based on Concepts
For each Cluster
    Perform link analysis on documents
    Form Macro-clusters based on domain
    Calculate Fuzzy totals within Macro-cluster
    Order Macro-clusters based on Fuzzy totals
End For
End If
Display Results

```

In the following example Macro clustering is applied to a very small sub set of the world wide web(3 web sites). Here, we assume that a search performed using the term "fuzzy logic" retrieves the results given in table 3.

Index	Document URL	Rel	FRF
1	www.cs.berkeley.edu/fz/internet.htm	15	1.00
2	www.pdn.ac.lk/fl/index.htm	14	0.93
3	www.pdn.ac.lk/fl/fuz.htm	13	0.86
4	www.cs.berkeley.edu/fz/papers.htm	12	0.80
5	www.pdn.ac.lk/fl/defuz.htm	11	0.73
6	www.fuzzylogic.com/index.htm	10	0.66
7	www.fuzzylogic.com/tech/index.htm	9	0.60
8	www.cs.berkeley.edu/fz/zadeh.htm	8	0.53
9	www.fuzzylogic.com/tech/papers.htm	7	0.46
10	www.pdn.ac.lk/fl/research.htm	6	0.40
11	www.fuzzylogic.com/tech/functions.htm	5	0.33
12	www.pdn.ac.lk/fl/new.htm	4	0.26
13	www.cs.berkeley.edu/basic.htm	3	0.20
14	www.fuzzylogic.com/extnsion.htm	2	0.13
15	www.fuzzylogic.com/neurofuzzy.htm	1	0.06

Rel. – Relevance value FRF – Fuzzy Relevance Factor

Table 3

For the sake of simplicity, we limit the area of the search to 3 pre-specified websites. The F.R. Factor is determined in relation to the order of the search results. (i.e. most relevant result appears first). Then we fuzzify the set of results with reference to the first result, which we consider to have complete membership in the fuzzy set $R = \{\text{documents that match the search query}\}$. The related function for the fuzzy set is given in fig.4

Degree of membership

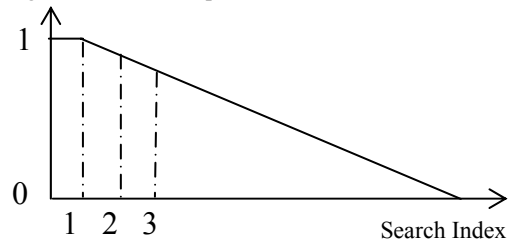


Fig 4 - Membership function of Fuzzy Set R

The results in the Table 3 are aggregated based on link names(web sites). The F.R.F. total for each domain is

also calculated, and the resulting table (Table -4) is sorted according to the F.R.F. total. Subsequently we obtain a list of websites related to the search term in order of relevance.

MC Index	Domain(Website)	FRF Total
1	www.pdn.ac.lk/fl/	3.18
2	www.cs.berkeley.edu/	2.53
3	www.fuzzylogic.com/	2.24

Table 3

The development of XML technologies for Internet applications has received attention in the industry as it's use give more meaning to the structure of a web document. XML is becoming a new standard of data representation and exchange, and more documents are expected to be available on the web. In XML, the structures and possibly the meaning of data are explicitly indicated by element tags. Therefore, incorporating XML will no doubt increase the accuracy of fuzzy content based information retrieval.

The concept 'Macro-clustering' was presented as a solution to improve search queries with a few search terms. The need for this arises because a majority of queries performed on search engines belong to this category.

7. Conclusion

The path to an 'perfect' web IR systems is still far ahead and the model for such a system is cannot be fixed as the World Wide Web changes at a rapid rate. Researchers have identified the use of soft computing methods(Fuzzy Logic, Neural Networks, Genetic Algorithms) as tools in making this goal a reality. This paper discussed the part played by Fuzzy Logic. Finally, the concept of "Macro-clustering" was introduced as a solution for optimizing results of generalized search queries. This concept contains many aspects on which further research can be carried out.

References

- [1] Sankar K. Pal and Pabitra Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, IEEE Transactions on Neural Networks, Vol. 13, No. 5, September 2002
- [2] C. V. Negotia, "On the notion of relevance in information retrieval.", Kybernetes, vol. 2, no. 3, pp. 161-165, 1973
- [3] Oren Zamir and Oren Etzioni, "Web Document Clustering: A Feasibility Demonstration"
- [4] Ellen L. Walker, "Image Retrieval on the Internet – How can Fuzzy Help?"
- [5] R. Tong , V. Askman and J. Cunningham, "RUBRIC an artificial intelligence approach to information retrieval". Proc. 1st. Int. Workshop on Expert Database Systems, 1984.
- [6] S. Miyamoto and K. Nakayama, "Fuzzy information retrieval based on a fuzzy pseudothesaurus", IEEE Tran. On Systems, Man, and Cybernetics, VOL. 16, 1986
- [7] Massimo Marchiori, "The limits of Web metadata, and beyond", MIT Laboratory for Computer Science, USA
- [8] R. Yager, "A framework for linguistic and hierarchical queries for document retrieval", Soft Computing in Information Retrieval Techniques and Applications, 2000, vol. 50
- [9] T. Geodeon and L. Koczy, "A model intelligent information retrieval using fuzzy tolerance relations based on hierarchical co-occurrences of words", Soft Computing in Information Retrieval Techniques and Applications, 2000, vol. 50
- [10] A. Molinari and G. Pasi, "A fuzzy representation of HTML documents for information retrieval systems", Proc. Of the 5th IEEE Int. Conf. on Fuzzy Systems, New Orleans, 1996
- [11] Mei Kobayashi and Takeda Koichi, "Information Retrieval on the Web", ACM Computing Surveys, vol. 32, June 2000
- [12] Mukherjean Sougata, Kyoji Hirata and Yoshinori hara, "AMORE: A World Wide Web image retrieval engine", World Wide Web, vol. 2, 1999
- [13] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separate Clusters", J. Cybernetics, vol. 3, 1973
- [14] R. Krishnapuram, A. Joshi, and L. Yi, "A fuzzy relative of the k -medoids algorithm with application to document and snippet clustering", Proc. IEEE Int. Conf. Fuzzy Syst., 1999
- [15] A. Joshi and R. Krishnapuram, "Robust fuzzy clustering methods to support web mining", Proc. Workshop in Data Mining and Knowledge Discovery, SIGMOD, 1998
- [16] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques with Java implementations", Chapter 6, Morgan Kaufmann Publishers, 2000
- [17] Timothy J. Ross, "Fuzzy Logic with Engineering Applications", McGraw-Hill, Inc., 1995
- [18] J. T. Yao and Y. Y. Yao, "Information Granulation for Web based Information Retrieval Support Systems"
- [19] Antonio de Padua Braga and Jason T. L. Wang, "Internet-based Intelligent Systems", International Journal of Computational Intelligence and Applications, Vol. 2, No. 3, 2002
- [20] R. Krishnapuram, A. Joshi and Olfa Nasraoui, "Low-complexity Fuzzy Relational Clustering Algorithms for Web Mining"
- [21] Naftali Tishby and Noam Slonim, "Data clustering by Markovian relaxation and the Information Bottleneck Method"
- [22] Thomas Minka "Design patterns – Proxy patterns", <http://vismod.www.media.mit.edu/~tpminka/>



Document Management Techniques & Technologies

Joseph P. Sathiadas¹, G.N. Wikramanayake²

1. Virtusa (Pvt) Ltd., Tel: 0777 313815, Fax: 074 724161, email: jpsathiadas@eureka.lk

2. University of Colombo School of Computing

Abstract

Electronic Document Management System (EDMS) is a rapidly developing technology and is considered as the solution for organizations that needs a way to manage the information efficiently. EDMS applications focus on the control of electronic documents throughout their entire life cycle, from creation to eventual archiving. Its functions include document creation, storage and retrieval, management, version control, workflow and multiple delivery formats.

Document management is not a single entity or technology, but rather a combination of elements. It is the use of information and different users in a business process, combined with the technology that permits the interaction. The technologies that make up the EDMS are categorized into distinct functional groupings. We present these and describes the techniques used to electronically manage documents. We also explores the immediate future of the EDMS and conclude that having the EDMS industry is at crossroads in its own lifecycle and is made up of a highly fragmented group of products with no single integrated vendor or framework for automating the entire cradle to grave document life cycle.

Keywords: Electronic Document Management System; EDMS; DMS; Resource Management

1. Introduction

Most of the organizations have vast amount of information that are required for their on-going projects or for their future projects in the form of knowledge of their workers or in documents. But, lack of information sharing among people and various project groups, lack of good management of information assets and lack of support from the knowledge workers make this information not available and not useful. Hence, the need for a system that could cater for this requirement and address these issues came up.

An Electronic Document Management System (EDMS) [4, 7] address most of these issues and is considered as the solution for organizations that needs a solution to manage the information efficiently. Although data management has been there for the last 30 years or so, document management came into the picture only about 10 years back. EDMS became popular with the advent of technology growth and computers.

2. Functions of EDMS

EDMS applications focus on the control of electronic documents throughout their entire life cycle, from creation to eventual archiving. Its functions [2, 6] include document creation, storage and retrieval, management, version control, workflow and multiple delivery formats.

Document creation: A document is a container, which brings together information from a variety of sources, in a number of formats, around a specific topic to meet the need of a particular individual or an organization.

Storage and retrieval: This involves storing and retrieving documents in a storage device such as, hard disk, tape etc.

Management: This covers a wide area of managing all the documents efficiently to cater to the needs of the organization and the individuals.

Version control: This is a way of tracking changes done to a document and the ability to retrieve old versions of a document.

Workflow: This is a way of tracking the state of the document and who is responsible for that step.

Multiple delivery formats: Ways of delivering the document content in various formats, such as PDF, Word, Image etc., to cater to the requirements of the end users.

3. The Document Management Space

Document management is not a single entity or technology, but rather a combination of elements. It is the use of information and different users in a business process, combined with the technology that permits the interaction. Hence, the Document Management Space can be divided into four major areas namely: documents, people, processes and technology.

Documents: The wealth of an organization is the information it has. Approximately, 20% of this information lies as data and the remaining 80% lies in documents. The 20% of data are normally well managed and maintained in databases. Lot of effort have gone into managing and utilizing this data, without giving much care about the documents that have most part of the information.

People: As like any other systems, document management system also serves a variety of different users. Users can be a Creator, Coordinator or a Consumer. A single person can also play multiple roles. The creator is the author and generates document content. Coordinators ensure that a document is properly reviewed and approved for release. They are responsible for assigning tasks for other members to perform on the document. They are responsible for the delivery of the document to the Consumer. Consumers are the real end users of the documents, who read or study them. Consumers rely on the Coordinator to get them what they want in an appropriate format.

Processes: When a document goes from conception to consumption, a process has to be in place to ensure that every thing is going as planned and according to the expectation. The figure 1 illustrates the process of converting Data into Knowledge.

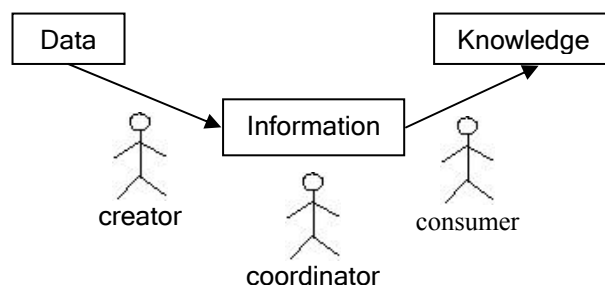


Figure 1: Converting Data into Knowledge

Technology: With the rapid technology growth, a powerful PC running a GUI is common in most of

the offices. Imaging, database, networks and desktop applications are some examples of the technologies associated with the Document Management Space.

4. EDMS Technologies

The technologies that make up the EDMS can be categorized into six distinct functional groupings namely: repository, conversion, indexing and searching, creation, workflow and distribution [8].

4.1. Repository

This is the place where documents/objects are persisted and restored for use. DBMS and/or file systems are the most commonly used repositories. When you consider a repository, it consists of a database and/or a file system as the backend, a server engine in the middle and a client system as the front end.

With the current operating systems optimized to handle files, most repositories are built with both a file system and a database. While the file system is used to store the actual document or object, the database is used to store the Meta data and rules. The database engine or a complete separate application can act as the server application. By the introduction of the server in the middle layer, the client becomes thin. The client system can be a stand-alone application. Now most of the systems provide a web based thin client as their front end.

Library Services (Check in/Check out services, Verification of privileges, Access Control), Version Control (Version numbering, Linear or Branched versioning, Creating and controlling versions, Archiving) and Configuration Management (Virtual Documents) are the primary functions provided by the repositories. The secondary functions provided by the repositories are conversion, searching and indexing, and workflow.

Open document management API (ODMA) and Document management alliance (DMA) are two main repository standards. ODMA is working towards a set of interfaces between desktop applications and document management client software and DMA is working towards a set of interfaces between document management servers.

4.2. Conversion

Most of the documents are comprised of text, images and multimedia objects. Since, each of these formats are fundamentally different, they have to be converted using different tools and techniques.

Conversion is generally regarded as a non-value adding process. But, conversion that helps to improve the performance of searching and retrieving is considered as value adding process.

Some of the common standard formats are:

Text – ASCII, SGML, HTML
 Graphics – CGM, IGES, TIFF, GIF, JPEG
 Multimedia – MPEG

Document conversion process is described by figure 2. Here, the source format is the format of the original documents and the target format is the format required/preferred by the end users. A filter is used in the middle to find the best way to do this conversion.

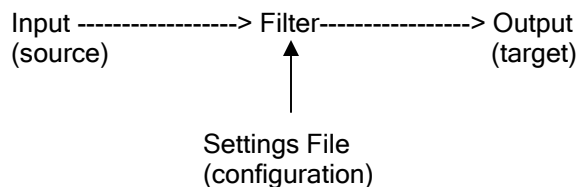


Figure 2: Document conversion

4.3. Indexing and Searching

4.3.1. Indexing

With the growth of the document repositories and online document sets, the speed of information retrieval becomes critical. Indexing allows a way of this by breaking up a document into more granular down to word level. Inversion of terms is one popular way of indexing. This method will have a sorted list of all the keywords in a file with pointers pointing to the actual location.

Normally the documents are fed to the indexing process initially. Some of the most popular indexing engines such as Open Text, Verity, Fulcrum [2] can accept most major word processing documents.

4.3.2. Searching

Any EDMS should provide the searching facility to the user. The user must be able to search for a particular document or a document containing particular information without delay.

Information retrieval effectiveness is measured by recall and precision. Recall is the proportion of the relative materials retrieved and precision is the proportion of retrieved material that are relevant.

The following types of search mechanisms are used.

Secondary search: Search within the results of the first search.

Semantics: Identify the exact definition of the word and then start searching.

Synonym: When searching for a particular word, search also for its synonyms (e.g. searching for Book or Publication).

Boolean: is the ability to merge together multiple words with operators like AND or OR.

Proximity: is the ability to search for multiple words grouped together.

Fuzzy word: is used when the word is not perfect due to some reason (OCR reading etc.). This uses character and string pattern searching to determine the correct word.

Concept: is the ability to pass information or news related to a concept.

4.4. Creation

Initially, document creation focused on creating paper output only. But, with the introduction of online document management system, features like document content, format and electronic structure also started to have an impact.

The following has to be considered when an online document is created:

- What the reader sees
- Hyperlink information
- Visual conventions
- Document sizing
- Number of files needed for online manual
- Purpose and content of information
- How the information will be used

Many fail to realize that documents are not mere static collection of texts, but are capable of embedding organizational knowledge and transmitting across the organization through the use of tools.

We could find three types of knowledge in documents, namely: structural knowledge (how a document is constructed), domain-specific knowledge (what the document contains) and contextual knowledge (how document sections relate).

4.5. Workflow

Workflow is the movement of a document through a series of steps to achieve a desired business objective. Workflow seeks to eliminate wasted time such as the time documents spend sitting in an in-box, the time taken to gather information to take action and time spent in moving documents from one person to the next.

Though workflow is not part of document management, having this feature enrich the functionality of the document management process. A document management repository controls documents, while a workflow engine controls the review and approval process.

For a document management system to give a meaningful solution a workflow must integrate properly with users, security, versioning, attributes, documents and relationships.

A workflow comprise of 4 primary elements, namely: process, actions, people and document.

Process: the sequence of steps necessary to reach an end objective, the business process. The two fundamental types of process are structured and ad-hoc.

Actions: what's to be done at each step? The two classes of workflow used as part of an EDMS are Review & Approval and Other tasks.

People: who is to accomplish these items? Normally this is specified by roles rather than by individuals.

Document: the focus of the process.

The key to implementing the workflow is not technical, but rather the people and their resistance to change.

There are four types of workflows, namely: sequential, parallel, branching and time drive.

Sequential: Linear set of steps. Each step is dependent on the completion of it's previous step

Parallel: Document can be passed to multiple people for action at the same time. This introduces an issue of reconciling the results of each parallel path.

Branching: Is a conditional type of workflow where paths to be taken are chosen based on a criteria.

Time driven: A time period is defined for a step, along with a action to be performed if the time period is exceeded.

4.6. Distribution

Distribution is the act of delivering the needed information, usually in document form, to the end user who will use or consume it. The format of the document will vary based on the organizational style and user needs.

Although many think that electronic document is the mode of distribution when it comes to EDMS, it is actually false. Many readers still prefer paper documents than electronic documents. Hence, EDMS should continue to support output in paper form. Nevertheless, electronic document delivery has many advantages over the traditional paper delivery. They are:

- Lower cost of document distribution
- Easier maintenance and updates to the documents
- Faster access to documents
- Greater nonlinear access to information
- Customized views of the document set
- Better quality of presentation

Distribution of documents takes place in four different ways as online publishing, offline publishing, repository viewing and web access. In the typical online publishing, documents are viewable by the consumer. In the offline publishing method, the documents extracted by the repository is published for use by the consumer. In the repository-viewing method the consumer searches and retrieves documents from the repository. Finally in the web access method (figure 3) the consumer uses a web browser to search and retrieve documents.

Document viewing is done for electronic paper, online documents and native formats. Tools are available for viewing of such documents. Important features of these tools are searching, zooming, hyper linking, annotations, outlining/tables of contents, bookmarks printing and integration capabilities.

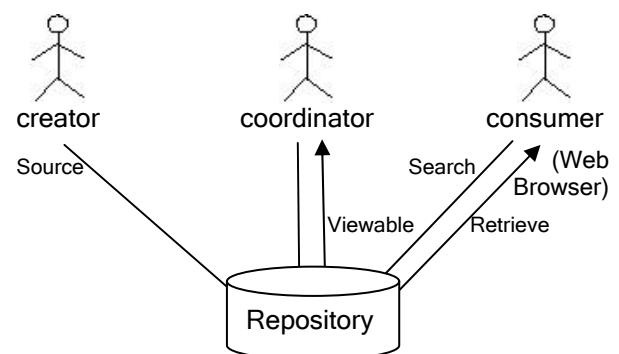


Figure 3: Web Access Method

Currently most of the document management tools have a 2D interface for the user to display the grouping of the documents. Good example is the Explorer of windows, which has a tree structure to display the directories and within it, it displays the document titles or thumbnails.

One interesting thought towards this is to explore the viability of providing a 3D interface for this purpose. The 3D interface can have a 3D plane, and the thumbnails of the documents can be arranged on this plane by the user. This design is expected to exploit the humans' natural capacity for spatial memory and cognition. The user has the advantage of placing the important documents close to him and the less important documents away from him. Finding a document is just like locating something in a house or room. Some amount of research has taken place in analyzing the performance and user preference of this approach. Research done by Andy Cockburn and Bruce McKenzie shows that 2D interfaces performed slightly better than 3D when it comes to storage and retrieval times. But, the users preferred 3D interfaces than 2D [3].

5. Techniques

The eventual goal of document management is a single electronic system combining paper and digital information [10]. The document management domain is continuously undergoing changes exploring new techniques to cater to the demand of the industry needs. As for the immediate future of document management is concerned, it can be said that the focus will be mainly on areas such as web growth, file compatibility, structured documents, efficient document storage and graphical interface. Structured documents, XML databases, file manager interfaces and intelligent documents are new technologies used for the management of documents.

5.1. Resource Management Design Pattern

The main intention of the Resource Management design pattern is to manage multiple resources of the same type. To maintain the status of managed resources, a Resource Manager is implemented having lists. Resource Managers are commonly used to control access to any class of "sensitive" resource objects. Resource Managers maintain two lists of resources, i.e. free list and busy list. Free list contains all the resources that are free. Busy list contains all the resources that are busy. A resource allocation request is serviced by the resource manager by allocating a resource from the free list and putting it

in the busy list. Similarly, a resource release request is serviced by the resource manager by inserting the freed resource to the free list.

A DMS can be considered as an instance of the Resource Management design pattern. In this particular instance, the documents are the managed resources and the Document Manager can be considered as the resource manager. Hence, the Document Manager allocates, tracks, controls, and de-allocates document objects. Clients cannot access documents directly. Instead, they must access these objects indirectly, through the Document Manager, using identification numbers provided by the Document Manager. In this way, illegal client requests can be detected.

Document Manager maintains free list and the busy list. When a document is requested by a client, the document manager first checks if that document exists in the free list and if it exists there, it retrieves the document and passes it to the requesting client. It removes the entry of this document from the free list and places it in the busy list. While maintaining the free/busy status of the documents, the Document Manager must be able to keep track of some additional state attributes of the document also. These can be facilitated by storing these attributes as a collection in the free/busy list along with the document information.

5.2. Structured Documents

While HTML has become the universal electronic delivery encoding of a document, it is neither presentationally rich enough to support high quality paper delivery, nor expressive enough to allow easy automatic transformations to multiple formats or to recombine the content. Generic structured markup like that provided by SGML or XML, opens many new possibilities for Document Management Systems (DMS's). A DMS that supports such standards natively is referred as a *Structured DMS*. The characteristics of a Structured DMS are as follows [1]:

'On-the-fly' creation of renditions: SGML and XML support delivery from single source in multiple formats. This is achieved by having a style sheet or filter for a class of documents. Instead of storing all these renditions, these can be created 'on-the-fly' as per the need.

Automatic transformations: This is the process of reordering document components or incorporating into new documents. For instance, entries from a dictionary identified as being legal terms might be

extracted from a general dictionary to create a new legal dictionary.

Access control at element level: This transformation capability can be used to strip out certain components from documents before they are delivered to a user.

Access to elements (component versioning): Most of the DMS's manage complete documents. Component versioning is the approach of versioning the components of a document.

Intensional versioning: Allows a number of rules to be applied to a version set to derive a particular *variant*.

Human-readable description of changes: Because the logical structure of documents is explicit in structured documents, differences between versions of a given document (commonly termed *deltas*) can be represented in terms of operations on its elements-*-syntactic differencing*.

Extended search capabilities: The markup in structured documents can be used for more sophisticated searching capabilities. Documents and elements can be retrieved based on structural relationships and a mixture of content and structure.

Document-based workflow: SGML and XML also provide a convenient syntax for describing data associated with business processes. Workflow *process definitions* (defining the sequence of activities that make up a business process) and *process instances* (recording the status of a particular business process in progress) can be represented as one or more structured documents. This facilitates mobility of the business process between systems and allows the process to be advanced between connections to a central server. Because these processes are just another SGML or XML document, they can be versioned like any other document in the DMS.

5.3. Storage: XML Database

The shift from SGML to XML has created new demands for managing structured documents. Many XML documents will be transient representations for the purpose of data exchange between different types of applications, but there will also be a need for effective means to manage persistent XML data as a database. The paper, *Requirements for XML Document Database Systems* [9], explores requirements for an XML database management system. The paper does not suggest a single type of

system covering all necessary features. Instead it aims to initiate discussion of the requirements arising from document collections, to offer a context in which to evaluate current and future solutions, and to encourage the development of proper models and systems for XML database management.

5.4. File Manager Interface

Most part of the time is spent looking for information than actually using it. This problem is the result of the shortcomings of the modern desktop. File manager software is no longer an effective tool for managing documents. Tools for creation and information exploration are disintegrated. Key contextual information is hidden from the user (we call this *the hidden web*). Search tools are impersonal.

Conventional file managers organize the documents based on the directory hierarchy. The computer directory tree is one of the oldest artifacts of the pre-web era and is virtually unchanged since its creation. After nearly 30 years, the only significant advancement in file management software is the overlay of a graphical interface on what is still a text-based directory.

The directory structure is a poor way for a human to organize documents, since we organize *contextually* as well as hierarchically. This problem is particularly apparent when documents contain numerous references, both to other user documents and to documents on the World Wide Web. Directories were simply never intended to highlight and manage the relationships between information *within* documents.

The paper, *The Personal Web* [11], describes the nature of the modern personal information space (*the personal web*) and a tool that improves on conventional file management for organizing and exploring that space. It is based on the concept that a user's web experience should be as personal as possible, flowing easily between user and web documents, following various types of document relationship "links", and involving searches that take into account *who* is doing the searching.

Studies are underway to explore the possibility of providing 3D interfaces for the file manager. The main idea behind this is to give the user the feeling that he is placing a document in a physical location. Humans have a tendency to remember a position, which is similar to a position he deals with in his daily life rather than a hierarchical arrangement of files.

5.5. Intelligent Documents

An *intelligent document* contains knowledge about itself and its environment. It supports assembly of documents based on inputs given by the user [5]. An *active* intelligent document is able to construct and transform itself dynamically.

One of the basic problems in document management is to provide on-demand generation of individualized documents through dynamic *document assembly*. Document assembly composes new documents from an existing collection of documents. Naturally, document markup and structure contribute to the retrieval of the document fragments.

Automated assembly consists of three phases [5], namely: The user expresses his demands, Appropriate documents or document fragments are found and returned and The returned fragments are merged into a single uniform assembled document. Hence, the final document will be composed with information from various different documents.

5.6. Industry Standards for EDMS

Although ODMA and DMA have been the industry standards for Document Management for the past several years, most vendors are now looking into XML to become a standard for document management. XML promises to succeed where SGML failed, by being easier to implement. It is expected that, apart from virtually universal support, XML will also offer for the first time the opportunity to embed metadata intelligence *within* the documents themselves. To gain the maximum benefits of XML, all the documents should be converted to XML. But, this will be an expensive operation when we consider the millions of legacy documents stored in different vendor specific formats. But, XML can be effectively used to store meta data of the documents. In DM, the documents are generally stored in the file systems as flat files, but the meta data is usually stored in the database for frequent quick access. Indexing and searching capabilities are provided through this meta data. Currently, most of the vendors have their own vendor specific meta data format in XML without using a industry standard meta data format. This is mainly due to the fact that a meta data standard in XML doesn't exist currently.

If a general detailed meta data format is defined in XML and if the industry accepts to use that as the standard to store meta data, then all the individual DMS can be considered as one large virtual DMS. Hence, search and index operations can be generalized. Any client will be able to look up the

information in a DMS provided they have the proper security credentials to do the operation.

Since the current trend is to web enable all the applications, DMS is also not an exception. But the industry standards that exist for DMS, such as ODMA, are not optimized for web and multiple platforms. A general, industry standard, robust framework is needed for DMS so that any web based client will be able to connect to the DMS server to obtain services. Web services is one technology that can be seriously considered for this.

When a client request for a service from the DMS server, it has to pass the necessary information for this service and DMS should be able to return the results after the service is performed. For this, the client and the DMS server have to use a common language for communication. Due to XML's flexibility and robustness, it will be one of the ideal candidates for this.

6. Conclusion

To conclude the paper we first summarise the benefits of EDMS and then address the current status of EDMS and its drawbacks.

6.1. Benefits of EDMS

The benefits of EDMS can be described as follows:

- Lower cost of document creation and distribution
No material (paper) cost involved during the creation phase, and the documents can be distributed as a softcopy to whoever concerned without the need to make hard copies
- Improved, customized access to documents
Users will be able to view all the available documents and open any of them by just clicking the mouse. Read, write permissions could be set per user by the owner or administrator
- Faster document creation and update processes
- Increased reuse and leverage of existing information
Information in a document could be reused and leveraged by giving links in other documents or by the concept of virtual documents
- Better employee collaboration
Inputs from all concerned could be accumulated in a single document in real time and the necessary changes could be done
- Reduced cycle times in document centred processes
- More complete regulatory compliance
- Refined managerial control and reporting
- Enhanced document control and security

The administrator could control documents by setting privileges and access restrictions.

- Improved productivity/Reduced headcount
- Better customer/Client satisfaction
- Quick and easy access

6.2. Current State

The EDMS industry is at crossroads in its own lifecycle. The industry is made up of a highly fragmented group of products with no single integrated vendor or framework for automating the entire cradle to grave document life cycle. Enterprises are currently trying to overcome this issue on an ad-hoc basis with no clear vision or path to the future for solving the complete document lifecycle problem.

6.3. Drawbacks of EDMS

Although document management has many advantages and benefits, it also has its risks and drawbacks.

Problems with the current state of the document management are as follows:

- Technologies are too difficult and take too much time to implement. In some cases the solution takes longer to implement than the life span of the technologies.
- Lack of standardization. Solutions involving documents are usually not compatible with one another
- Pseudo-standards have emerged that are still vendor specific.
- Difficulty in managing documents independent of the application.
- Typical document solutions are implemented in phases. The technology to create, manage, and archive these documents must be as modular as the implementations.
- The idea of plug-and-play has never been implemented past most marketing departments.
- Integrators have limited resources to learn new tools.

When there exist ways to publish documents online, many organizations rush headlong and put the documents online without proper document creation and management processes in place. Online distribution without proper document control tends to create problems, since it could lead to wrong information or outdated information being sent to a large group by mistake. A well-designed EDMS should address this problem effectively, e.g. date stamps each document.

Another drawback with EDMS is that it is driven by documents and technology, but not the end users. Hence, the focus on the end user is lost and this tends to give a negative impact on the overall system. Document create process can address this problem by making the end user contribute towards it.

Most of the organizations, when provided with an EDMS, tend to dive in and publish all their documents, irrespective of their state, quality and the need. This leads to information overload. Hence, a process should be in place for document control and this should actively involve the authors and users. Finally, the end users, especially the new users, might react negatively to the introduction of the system because the learning curve is large and they are used to the less challenging print based media.

References

1. Arnold-Moore Timothy, Fuller Michael, Sacks-Davis Ron, "System Architectures for Structured Document Data", Markup Languages Vol. 2 No. 1, 2000, pg. 11-39, accessed on 101103, <http://www.mds.rmit.edu.au/~msf/papers/MT99.html>
2. Bielawski Larry and Boyle Jim, "Electronic Document Management Systems: A user centered approach for creating, distributing and managing online publications", Prentice Hall Computer Books, November 1996.
3. Cockburn A. and McKenzie B. "3D or Not 3D? Evaluating the Effect of the Third Dimension in a Document Management System", Addison-Wesley, Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems, Seattle, Washington, March 31-April 6 2001, pages 434-441, accessed on 1011003, <http://www.cosc.canterbury.ac.nz/~andy/papers/chi01DM.pdf>
4. Condon Thomas A., Roberts Doug, Nash Dawn, "Understanding EDMS: A guide to efficiently storing, managing, and processing your organisations documentation", white paper, 2002, accessed on 101103, <http://www.sdichicago.com/insidetheitstudio2002Q2/edmsfeature.pdf>
5. Document Management Research Group, Department of Computer Science, University of Helsinki: Structured and Intelligent Documents, accessed on 101103, <http://www.cs.helsinki.fi/research/rati/sid.html>
6. Functional Assessment of Hummingbird Enterprise, July 2002, accessed on 101003, <http://devx.newmediary.com/abstract.aspx?&scid=231&docid=37933>

7. Johns Hopkins Center for Information Services: Document Management Systems Recommendations, June 2002, accessed on 101103, <http://it.jhu.edu/divisions/nts/status/systemsrec.html>
8. Meyers Scott and Jones Jason, "Document design for effective electronic publication", Proceedings of the 5th Conference on Human Factors & the Web, June 1999, accessed on 101103, <http://zing.ncsl.nist.gov/hfweb/proceedings/meyers-jones/>
9. Salminen A. and Tompa F.W., "Requirements for XML document database systems". In E.V. Munson (Ed.), Proceedings of the ACM Symposium on Document Engineering (DocEng '01) (pp. 85-94). New York: ACM Press, 2001, accessed on 101103, <http://www.cs.jyu.fi/~airi/presentations/XMLdatabases.ppt>
10. Smith Brady, "The Future of Document Management", February 2002, accessed on 101103, <http://www.arches.uga.edu/~cpeter/Future.htm>
11. Wolber David, Kepe Michael, Ranitovic Igor, "Exposing Document Context in the Personal Web", Proceedings of the International Conference on Intelligent User Interfaces (IUI 2002), San Francisco, CA, 2002, accessed on 101103, <http://www.usfca.edu/~wolberd/papers/iui2002Final.pdf>



Public Key Infrastructure Security and Interoperability Testing and Evaluation

Job Asheri Chaula
Stockholm University/KTH
Forum 100, 164 40 Kista, Sweden

E-mail: si-jac@dsv.su.se

Louise Yngström, Stewart Kowalski
Stockholm University/KTH
Forum 100, 164 40 Kista, Sweden

E-mail: louise@dsv.su.se

Abstract

Public Key Infrastructures (PKIs) are currently being deployed in increasing sizes, numbers, fast changing technologies, and varying environments but our operational experience to date has been limited to a relatively small scale and small number of environments. Consequently, some open technical and environmental interoperability problems about the ways in which PKIs will be organized and operated in large-scale applications need to be addressed. For instance, (1) Non interoperable proprietary vendor-provided public key infrastructures (2) the distribution of revocation information which has serious security implications and the disadvantage to be very costly when running large scale PKI. This paper introduces the concept of security testing and evaluation to maximize PKI application security as a basis for PKI systems interoperability.

Keywords: Interoperability, Testing, Evaluation, Functional requirements, Target of evaluation

1. Introduction

Public Key Infrastructure is a system that is used to communicate securely and with trust among entities in a closed, distributed or open distributed computing environments. The trust amongst users is achieved through certificate exchange and authentication. Security in PKI is achieved through encryption of data. PKI is a particularly powerful information protection tool for systems and services on the Internet [23]. Currently, so many individual organizations, small, medium and large also incline to using PKI as a

solution to their internal information security problem. However, large-scale implementation of PKI has numerous interoperability problems. Problems related to non-compliance and non-scaling are regarded as technical problems while those which are related to policy, legal issues, and privacy are regarded as non technical problems.

Currently, at the centre of efforts to improve security are a group of security protocols such as Secure Electronic Transaction (SET), S/MIME, IPsec and TLS. All these protocols rely on public key cryptography to provide security services such as confidentiality, integrity, authentication, non-repudiation and Access control. Access control security service can be achieved through the use of attribute certificates.

PKI is responsible for binding public keys into certificates and managing those certificates in their life cycle. The part of PKI that is responsible with generation, issuance, and revocation of certificates is referred to as Certificate Issuing and Management System or CIMS. [18] defines X.509 version 3 certificate and the basic PKI components. The basic PKI components are:

The certification authority (CA): Issues and revokes certificates

Registration authority (RA): Vouches for the binding between public keys and certificate holder identity and other attributes

Certificate holder: Users who make use of certificates to encrypt/decrypt information.

Clients: Validate digital signature or encrypted messages and their certificate chains from the root CA

Repositories: Stores and makes available to users certificates and Certificate Revocation Lists. This

paper describes test assertions that are necessary to achieve secure PKI interoperability.

2. Large scale PKI Implementation Problems

2.1 Interoperability

Interoperability between PKI implementations forms the basis of security infrastructure. PKI infrastructure, just like telephone infrastructure, transportation system, water, power lines or gas supply system must therefore recognize the same fundamental doctrines and benefits to users. Users can be a group of users, an organization, society, nation or community of nations like the African Community, Asian community or the European Community. Unfortunately, PKI so far has not met this requirement due to manufacturers having non-interoperable products. PKI must therefore be accessible by every application, object, or any other entity that requires the security infrastructure's service. The cost of failing to achieve this is for example lost opportunity because of the failure to achieve communities of trust between borders, and more importantly to engage in legal electronic transaction [23].

2.2 CRL

The distribution of Certificate Revocation List (CRL) has the potential to be the most costly aspect of running a large-scale (PKI). In order to realize the cost effectiveness, several alternatives of revocation distribution mechanisms have been proposed. Some of the proposed certificate revocation mechanisms includes the On line certificate status protocol (OCSP), Delta certificate revocation list [18], complete or base CRL, sliding window delta CRL, authority revocation list, Freshest CRL, Redirect CRL, Dynamic CRL distribution point, Indirect CRL, etc.

Large scale PKI will involve voluminous growing numbers of users accessing the repository at the same time; in these scenarios issues of scalability, trust, request rate, reliability and timeliness are of great importance. For this reason implications of certificate revocation mechanism to end-user applications are to be considered when choosing CRL mechanism. However, the choice of CRL mechanisms may differ in different environments.

In order for a certificate to be valid, the certificate user must trust the issuing party utilizing any of the revocation status verification mechanisms. In order to maintain this environment of trust it is vital that the revocation process is well defined, implemented, and enforced without any ambiguity existing as to the

status of a certificate. This has serious security implications especially if the certificate is used to access organisation's sensitive data. If the CRL mechanism is compromised a malicious user whose certificate validity period has expired may gain access to the system.

2.3 Standards

Today, users can choose among a wide range of PKI standards that are available from different vendors. Most PKIs are built around the traditional X.509 standard, the Internet standard PKIX and the Public Key Cryptography from RSA PKCS to name just a few. However, this poses serious problems as the number of differing infrastructure implementations increases as components are added to it. This has the disadvantage that the implementation and support cost increases and interoperability becomes even more difficult.

2.4 Trust

There are several models or architectures that PKIs can employ to provide trust. One of the traditional widely used model is the hierarchical model [23]. This model is used in the [RFC 3280] standard. In this model users place the trust on a single CA referred to as Top Certification Authority who issues certificates to subordinate CAs in a chain. Other trust models are the web of trust mesh [23]. These are commonly used in e-mail programs and web browsers. However, most of the contemporary trust model work fine if they are deployed in a single organisation community of users. For multiple domains the questions of top CA of one domain trusting the top CA of another country is usually not easy to answer. This problem has been dealt with cross certification authority and Bridge CA models solutions and will work well for organisation and nations that mutually agree to do e-commerce and transfer electronic information in a secure manner. It is difficult to achieve a large scale web of trust (for example say a global web of trust) simply because not all countries or organisations trust each other, this is due to political differences, cultural differences and even economic differences.

2.5 Certificate Chain Verification

A certification path is a systematic ordered list of certificates starting with a certificate issued by the relying party's trust root, the TCA, and ending with the target certificate that needs to be validated the end entity (EE) certificate. Certification path validation procedures are based on the algorithm supplied in ITU-

T Recommendation X.509 and further defined in Internet Engineering Task Force (IETF) Request for Comments [18]. Certification path processing verifies the binding between the subject distinguished name and/or subject alternative name and the subject public key defined in the target certificate. The binding is limited by constraints, which are specified in the certificates that comprise the path, and inputs that are specified by the relying party. To ensure secure interoperation of PKI-enabled applications, the path validation must be done in accordance with the X.509 standard i.e. the RFC 3280 specifications [18].

2.6 Privacy

Privacy, as defined by Westin, “is the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others”[24] is a social and legal issue that has for a long time has drawn the attention of social scientists, philosophers, and lawyers. With the arrival of the computer and increasing capabilities of modern IT-systems and communication networks, individual, government and organisations privacy is increasingly endangered. Especially on regional conflicts, terrorism, business competition and nations plans to erect information highways (for example the Africa 1 project), there are severe privacy risks. Privacy as a fundamental human right has to be protected. Now, as we have discussed earlier, large scale PKI implementation involves different community of users from different organizations and possibly different countries.

Hubner states that: “Privacy as a fundamental human right has to be protected” [25]. As we have discussed earlier, large scale PKI implementation involves different community of users from different organizations and possibly different countries. All the PKI participants will have to use certificates with credentials that will be used to independently verify the binding of the name and the individual or organization taking part in the transaction. The private information in the certificate includes subject name, extension that possibly includes alternative name, postal address, picture, position in the company, country, zip, age and gender [18]. These details are too sensitive to let them be in a third party possession. Exposure of e-mail address to unintended people has the potential to become a target of spam e-mails.

2.7 Security Concepts

Security services are designated to protect organisation’s assets against various threats. Organisations seek security systems that provide one or

Large scale PKI will involve a great number of users accessing the repository at the same time. In these scenarios issues of request rate, reliability and timeliness are of great importance. For this reason implications of any trust model and certificate revocation mechanism to end-user applications require more research. Even though the revocation process is well defined, implemented, and enforced without any ambiguity existing as to the status of a certificate, in a large scale PKI implementation chain certificate verification will cause network bottleneck which eventually degrades the whole system availability. more security services. However, understanding the assurance level of the security system requires system testing that is based on established standards like the Common Criteria [7]. The concept of security organisation and the organisation’s assets is depicted in figure 1. Understanding the concept of system’s security testing is so important for organisations which IT infrastructure is central to the organisation’s well being.

Security policy is a statement outlining the organisation’s commitment to securing its assets. The countermeasures, vulnerability assessment, implementation of a procedure, systems testing, accreditation and certification will be performed in accordance with the organisations policy. Procedures and mechanisms have to be verified simply because a faulty procedure or mechanism leaves residual vulnerability that can be exploited to cause residual risk to the assets [7].

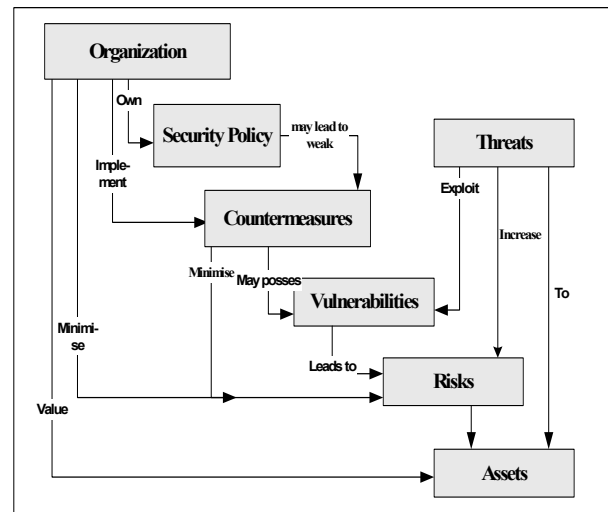


Figure 1. Security concepts [7 Page 14]

A threat to an IT asset is a potential violation of security mechanism [12]. Matt Bishop outlines the

following threats: Snooping, modification or alteration, masquerading or spoofing, repudiation of origin, denial of receipt, delay, and denial of service. Proper mechanisms to protect assets against these threats a security policy has to be in place. To stop different attacks appropriate security mechanisms have to be put in place to provide different security services.

Threats can cause damage to the IT assets which organizations have placed values. Threats can cause the IT assets to permanently or temporary stop performing its function or perform at a lower standard examine IT systems testing, security metrics, process metrics, organisations IT security policy, and implementation of security mechanisms the ultimate purpose is to have effective security services in place.

3. Security testing and verification

There are many misunderstanding related to understanding and describing security functional testing, validation, assessment, verification and evaluation of IT security functionality. The difficulty is partly due to issues of testing coverage, environmental issues (i.e in which environment is testing done), independent testing, assurance levels and threat analysis and lack of understanding of security metrics. Schneir [2] underlines that normal security testing fails because first, security flaws can appear anywhere either in trust model, system design, the algorithm and protocol, the implementation, the source code, the human computer interface, the procedure and the underlying computer system (hardware or system software). A single flaw can break the security of the entire system.

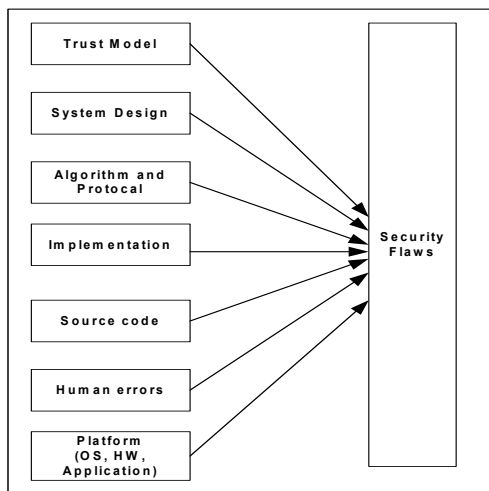


Figure 2. Security flow concepts

or capacity. This is referred to Loss of availability [7], unauthorized entities gaining access to the IT asset is referred to loss of confidentiality [7], an entity having unauthorized modification of an IT asset is referred to loss of integrity [7], an entity denying having acted on IT asset is referred to nonrepudiation and an entity having gained access to unauthorized IT assets is referred as to loss of access control. Understanding of these terminologies is one of the central issues in IT security because when we

Security testing is different from normal software testing practice. This is because security is a chain and dependant on the implementation, environment, user knowledge and only as secure as the weakest components. While system patching may have minimal effects on the functionality of the product, this may have severe impact on the functionality of security [2]. Software functional testing works well in software testing but in security functional testing does not reveal flaws until a security-testing expert systematically looked into it [2]. The only way to have confidence over any system security is to overtime, have expert evaluate it using established methods like the Common Criteria [2].

3.1 The role of a Security Testing Lab

Security testing lab is a networking environment designed for conformance testing, evaluating exploits, penetration testing, and performing various simulations. In same case a lab may provides a facility for hands-on training, practice, and exercises for testers to gain deep understanding of what is to be done in real life situations. The security of the physical, network and personnel security of the lab itself is very important to protect the evidence and the test results [26]

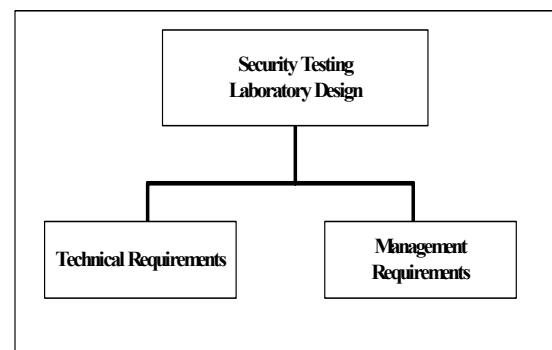


Figure 3. Security laboratory design

3.2 Management Requirements

The lab must be organized so that it can carry out its work according to the laid out procedures to avoid conflicts between developers and testers. The lab must have management and technical personnel with the authority and resources needed to do the testing. The lab must have policies to protect the clients' confidential information and they must have an organizational chart that specifies the interrelationships of all personnel [26].

3.3 Evaluation evidence control

The lab must have procedures to control all documents that are part of its quality documentation [26]. The lab must keep a master list of issued quality system documents. Changes to documents shall be reviewed and approved by the same authority performing the original review.

3.4 Lab continuous quality

The lab shall have procedures for quality and technical records including security and confidence measures [26]. The lab shall periodically perform internal audits of its activities. The executive management of the laboratory shall periodically review the lab's operations to ensure continuing suitability and effectiveness.

3.5 Technical Requirements

Testing requires clearer understanding of the TOE. The technical personnel must be qualified for each specific test and they make the most important component of the lab [26]. This can be achieved through appropriate training, education, experience on using the methodologies, or demonstrated skills. The lab shall use appropriate test procedures and standard methods. The lab has to documented instructions on the use methods and equipments.

3.6 Reporting the Results.

The test results accuracy reporting process is essential. Therefore, it is important to be reported accurately, clearly, and unambiguously [26]. The common criteria testing lab (CCTL), [26], provides controlled environment where by the internet access in controlled and each room is physically secured and can be used as isolated independent lab [26]. This is essential to make sure the evaluation evidence and evaluation results are secured.

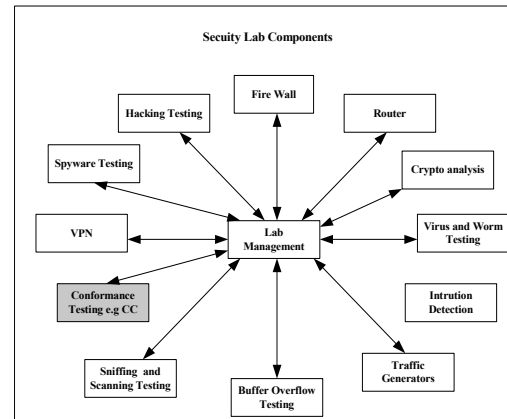
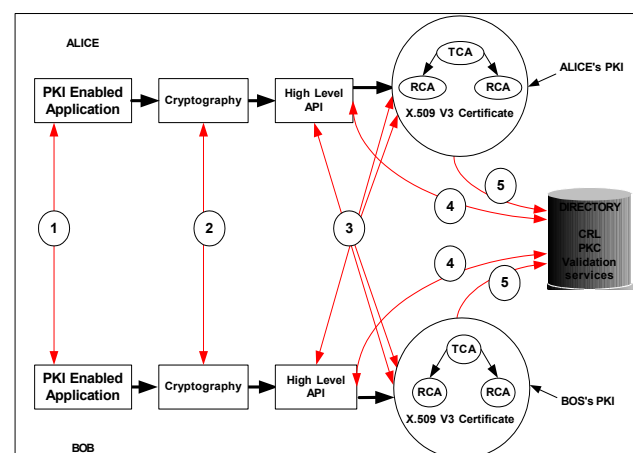


Figure 4. Security laboratory components

4 PKI Interoperability testing

The operational transaction between the two PKI enabled applications in Figure 5 exchanging information securely using the underlying cryptographic algorithms and the PKI. The application has one critical function of verifying the chain of the public key certificate to prove whether it can be trusted or not. The main purpose of our work in this area is to develop security functional testing assertions that can be used by users, developers and testers to verify the applications capability to validate public key certificates



4.1 Certificate Validation Criteria

In this section we provides the test assertion for testing path validation software against certificates defined in ITU-T Recommendation X.509 and further defined in Internet Engineering Task Force (IETF) Request for Comments (RFC 3280)[18]. Certificate path processing and includes determining that the certificate has been issued by a recognised trust anchor or its trusted subordinate, the digital signature of the certificate is valid, the certificate is within its stated validity period, the certificate has not been revoked, the certificate is being used in a manner which is consistent with its policy constraints, name constraints, and intended usage restrictions [18]

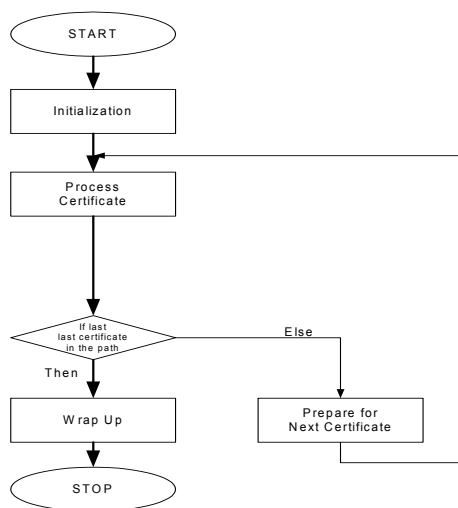


Figure 6. X.509 Certificate path validation sequence [18 Page 58]

Path processing initialization require that the following inputs and assumptions are provided:

Correct current time

The certification path length n where certificate n is the certificate to be validated and

The following must be always true if k is a certificate in the certification path

$\forall k$ element of $\{1, \dots, n-1\}$, the subject of certificate k is the issuer of certificate $k + 1$ [18]

Certificate 1 is the first in the certification chain and is issues by the TCA who is the trusted anchor in our case.

$\forall k$ element of $\{1, \dots, n-1\}$, the certificate is valid at the correct time in question[18]

$\forall k$ element of $\{1, \dots, n-1\}$, k does not include self issued certificates. This implies that self-issued certificates are not counted in the certification path. A certificate is a self issued certificate if the Distinguished Name (DN) that appear in the subject and issuer fields are identical and are not empty[18]

Policies that are acceptable to the certificate user are listed in evaluation evidence RFC 3280 [18]:

- Trusted anchor name, public key algorithm, trusted public key
- Initial policy mapping. This indicates if policy mapping is allowed
- Initial explicitly policy.
- Initial-any-policy-inhibit.

The certificate using application is the target of evaluation (TOE).

5 Security Functional testing

Security functional testing is applicable where some threats to security are viewed as serious [9]. Therefore, it is important to identify the threats prior to testing. The test assertion we are developing can also be used by where independent assurance is required to support contention that through testing has been done with respect to the protection of certificate attributes [7]. Security function testing is one important step to make sure the application can interoperate in mult PKI environment. In this testing it is recommended to make sure the applications ability to handle invalid certificates and valid certificate is tested.

6 Test cases

Each test case we have considered is selected based on the threat(s) we are mitigating. A summary of security services provided by each test case is provided in Table 1

6.1 Signature Test

Signature and names must initially chain, this means the signer of the current certificate in the chain must be the issuer of the next certificate in the chain. Dates must be correct, intermediate certificate include a basic constraint extension that asserts CA is TRUE, and the key usage extension asserts both keyCertSign and

cRLSign.[18]. Threats include infringement of data integrity, nonrepudiation, availability, and loss of confidentiality And authentication.

6.2 Validity period tests

Validity period is the time from the certificate was issued to the time the certificate is revoked. This is time interval that the CA warrants that it will maintain information status of the certificate until it is revoked or destroyed or archived. This interval is marked by validity beginning date (notBefore) and validity end date (notAfter). Both notBefore and notAfter can be encoded in UTC or GeneralizedTime and the certificate using application must support UTC and Generalized Time encoding [18]. Threats that are being addressed by this testing include unauthorised access to assets by terminated employees, business partners whose association does no longer exist, accessing resources before or after authorised time, loss of legal rights etc.

6.3 Subject and Issuer Name chaining test

PKI enabled application must verify if the issuer name in each certificate in the certification path matches the subject name of the preceding certificate in the path. Also the following parameters must be checked: Chaining order, Capitalization, unique identifiers, Mandatory attributes and Optional attributes [18]. Associated threats are infringement of data integrity, confidentiality and unauthorised access of assets.

6.4 Key Usage Tests

Key usage extension defines the purpose of the key contained in the certificate. This includes: digitalSignature, nonRepudiation, keyEncipherment, dataEncipherment, keyAgreement, keyCertSign, cRLSign, encipherOnly and decipherOnly[18]. The purpose of these tests is to determine the ability of PKI enabled application to process the key usage extension in the certificate. If the extension is labeled true (e.g. certSign true) it implies that the key can be used to sign certificates [18]. The threats associated to this testing are Unauthorized entity can issue a certificate that can eventually be used to gain unauthorized access to assets.

6.5 Name constraint test

The name constraint testings are designed to verify PKI enabled application ability to process the name

constraints extension. The tests in this section include certification paths in which one or more certificates include a name constraints extension with the following: DNS name, and uniform resource identifier (URI), rfc822 name and distinguished name (DN)[18]. Threats being address with this testing are: Denial of service and masquerading.

6.6 Policy test

The tests in this section are intended to verify if the PKI enabled application is able to process the certificate policies extension, including the ability to process policy qualifiers and the special policy object identifiers [18]. Threats that are addressed with this testing are: Misuse of access rights

6.7 Distribution point test

The tests in this section are designed to verify PKI enabled application ability to process the CRL Distribution points certificate extension and the issuing CRL extension [18]. These two extensions may be used for multiple purposes: to spread certificate status information about the certificates issued by a CA among multiple CRLs, to have certificates listed on different CRLs depending on the reason that they were revoked, or to have the CRL that indicates the status of a certificate be issued by a different entity from the CA that issued the certificate. In many of the tests in this section, the certification path includes a certificate for which there is no valid, up-to-date certificate status information available. For these tests, the application must either reject the certification path or warn the user that the status of the certificate cannot be determined [18]. Threats that can be hedged through this testing are: Committing transaction using revoked certificates.

6.8 Basic constraint test

The basic constraint testing are to be used to determine if an application properly processes the basicConstraints extension as specified in X.509[18]: If extension is present and is flagged critical, or is flagged non-critical but is recognized by the certificate-using system [18], then If the value of CA is not set to true then the certified public key shall not be used to verify a certificate signature; If the value of CA is set to true and pathLenConstraint is present then the certificate-using system shall check that the certification path being processed is consistent with the value of pathLenConstraint[18]

6.9 CRL Test

The PKI enabled application must be able to retrieve valid revocation data for each certificate in the path. The date should include the following: Revoked and missing certificates, CRL issuer name, extensions, serial number, next update and signature. If the application is unable to retrieve valid revocation data for one or more certificates in the path, it must reject the certification path. In the following tests, it is assumed that if an application is unable to find valid, up-to-date certificate revocation list for each certificate in the path, that either path validation will fail and the application will display a warning to the user indicating that the status of the certificate cannot be determined.

Threats that are being addressed by this testing are: Fraudulent and users with expired access right gaining access to the assets. Such users can be terminated employer, contractor whose contract period is expired, a user who has terminated association with a CA, A doctor who for some reason is no longer required to access patient records etc.

6.10 Private certificate extension

The tests in this section are designed to verify PKI enabled application ability to process certificates that include unknown extensions. Unknown extensions that are marked non-critical may be ignored whereas an application must reject a certificate that includes an unknown extension that is marked critical [18]. Threats that are being hedged by this testing are: Unauthorised use of a certificate to commit a transaction

6.11 Self signed test

CA's certificate signing keys are used for a limited periods of time for security reasons [18]. Typically, at the time that a CA rekeys, the previous public key is still considered to be valid. This allows relying parties to continue to validate certificates that were signed using the previous private signing key gracefully. In order facilitate continuity of operations, the CA, at the time of rekey, will issue two self-issued certificates [18]: one certificate that contains the new public key that is signed using the old private key and one certificate that contains the old public key that is signed using the new private key.

Threats that are being addressed by this testing are: Denial of service.

Test Cases	Threats related to Security Services					
	Integrity	Confidentiality	Availability	Nonrepudiation	Access Control	Authentication
Signature	X	X	X	X	X	X
Validity Period	X	X	X	X	X	X
Subject and issuer Names Chain	X	X	X	X	X	X
Key Usage	X	X	X	X	X	X
Name constraint test			X			X
Policy test	X	X	X	X	X	X
Distribution point test			X			
Basic constraint test	X	X		X	X	X
CRL Test	X	X		X	X	X
Private certificate extension	X	X		X	X	
Self signed	X	X	X			X

Table 1. Threat and test cases table

7 Conclusion

We have presented PKI enabled application security functional testing assertions that are necessary to address serious security breaches that can occur if the application cannot verify certificates correctly and continuously. In normal software testing faults are called bugs. Bugs are also minimized through testing and evaluation. In research we focus on bugs that not only have effect on the functionality of the software but also have serious security implications. This is central for the interoperability of PKI application in a mult PKI environment. Detailed work involves defining the tests for each sub headings in section 6 and the criteria for critical and non-critical tests. PKI interoperability testing further work involves testing metrics development that is useful for tracking coverage of the test assertions and finally be able to use empirical data to gage the security that can be achieved in different test cases. PKI testing is not easy to simulate in a lab because other interoperability problems are not technical but rather they are related to policy, privacy and legal issues and the environment. The testing we have presented in this paper is limited to the application's ability to verify X.509 certificates correctly.

References

- [1] William Stallings (2000), Network security essential applications and standard
- [2] Bruce Schneier, (2000), Secrets and lies: Digital security in a networked world
- [3] Dieter Gollmann (2000) Computer Security
- [4] ITSEC (1991), Information Technology Security Evaluation Criteria
- [5] CTCPEC (1993), Canadian Trusted Computer Product Evaluation Criteria (CTCPEC) version 3.0
- [6] CCIMB-99-031 (1999), Common Criteria for Information Technology Security Evaluation: Introduction and General
- [7] CCIMB-99-031, 1999, Common Criteria for Information Technology Security Evaluation: Security assurance requirements Version
- [8] CEM-99/045, 1999, Common Methodology for Information Technology Security Evaluation Methodology Version 2
- [9] SSE-CMM, 2003, Systems Security Engineering Capability maturity Model, (SSE-CMM Version 3) <http://www.sse-cmm.org/model/ssecmmv2final.pdf>
- [10] CMM, 1995, Systems Engineering Capability Maturity Model, (SE-CMM Version 1.1) <http://www.sei.cmu.edu/cmm/cmms/cmms.html>
- [11] SW-CMM, 2003, CBA IPI and SPA Appraisal Results 2002 Year End Update (CMM based Appraisals for Internal Process Improvement (CBA IPIs) and • Software Process Assessments (SPA)
- [12] Bishop 2002, Matt Bishop (2002) Computer Security Art and Science
- [13] CSRC, Development of High level PKI Service API” NIST, Available at <http://csrc.nist.gov/pki/pkiapi/welcome.htm>
- [14] PKIX, IETF, “Public Key Infrastructure” <http://www.ietf.org/html.charters/pkix-charter.html>
- [15] Stuart McClure, Joel Scambray, George Kurtz (2001) Hacking Exposed: Network Security Secrets & Solutions, Third Edition
- [16] Ross Anderson (2001) Security Engineering: A guide to building dependable distributed systems
- [17] Ed. Skoudes (2001) Counter Hack: A step by Step Guide to computer Attacks and effective defenses
- [18] RFC 3280, 2002, X.509 Version 3 Certificates and CRL version 2
- [19] RFC 1321, “The MD5 Message-Digest Algorithm”, R. Rivest, MIT Laboratory for Computer Science and RSA Data Security, Inc. April 1992. Available at: <http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1321.html>
- [20] PKC7, RSA Laboratories, “Cryptographic message standard” Public-Key Cryptography Standards <http://www.rsasecurity.com/rsalabs/pkcs/pkcs-7>
- [21] FIPS 180-1, “Secure Hash Standard” <http://csrc.nist.gov/cryptval/shs.html> (Read in October 2002)
- [22] Bruce Schneier 1993 “Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish)” Available at: <http://www.counterpane.com/bfsverlag.html>
- [23] Russ Housley (2001) Planning for PKI: Best Practices guide for deploying public key Infrastructure
- [24] Alan F. Westin, 1967, Intrusions on Privacy: Self-revelation, Curiosity, and Surveillance in: Alan F. Westin, *Privacy and freedom Atheneum, New York, 1967 pp. 52-63*
- [25] Simone Hubner, 2001, Privacy in the Global Information Society [Online] Available at: <http://springerlink.metapress.com/app/home/content.asp?wasp=87exf5a92g1trh9aaw7m&referrer=contribution&format=2&page=1> (Accessed in Sept. 2003)

- [26] Robert L. Williamson, Jr., Tammy S. Compton, James L. Arnold, Jr., and J. Mark Braga Science Applications International Corporation (SAIC) Common Criteria Testing Laboratory (CCTL) Technical Directorate, SAIC CCTL. Available at:
<http://www.saic.com/infosec/pdf/CCTL-ITEA.pdf>. (Read in March 2003)



Authorization System in Open Networks Based on Attribute Certificates

Jeffy Mwakalinga¹, Eric Rissanen², Sead Muftic³

¹Department of Computer and System Sciences, Royal Institute of Technology, Kista, Sweden

²Swedish Institute of Computer Science, Kista, Sweden

³Department of Computer and System Sciences, Royal Institute of Technology, Kista, Sweden

Computer Science Department, The George Washington University, Washington DC, USA
[jeffy@dsv.su.se, erik.rissanen@bredband.net, sead@gwu.edu]

Abstract

This paper describes a security system for authorization in open networks. Authorization means authority to access certain resources, to perform certain operations, or to use certain system functions. In this paper the authorization system is based on use of attribute certificates. An attribute certificate is a signed object containing authorization attributes of a user. Before checking whether a user is authorized to perform an action or to access an object, the identity of the user must be verified. The identity verification system is based on public key certificates. We separate authorization system from authentication system because the same authority does not always establish authorization and authentication information. However these two systems must be combined and that is done by including the serial number of the user's public key certificate as a field in the user's attribute certificate, which carries authorization information

The topology of the authorization system comprises authorization authority servers issuing attribute certificates to users, application clients handling those certificates, and application servers verifying user access rights based on attribute certificates. Furthermore, all these components are themselves certified by standard PKI certification authorities, thus supporting mutual authentication and cross-domain scaling.

Keywords: Certification authority, attribute certificate, attribute authority, authorization and access control models.

1 Introduction

1.1 General Principles

This paper describes a generic system for authorization in open networks based on attribute certificates. Authorization means authority to access certain resources, to perform certain operations, or to use certain system functions. Authorization addresses three major problems: identification of users and assignment of globally recognized roles; matching of user roles with authorization attributes like security labels; enforcement of authorization privileges and making decisions. Today organizations run Web servers and resources of these servers should be accessed globally only by authorized people. For instance, companies have IT resources, which may only be accessed by customers who subscribe for them. In most cases a customer, who subscribes for resources, is given a user name and a password and can log in using these tokens to the servers. A user should be able to access the resources from any machine in the global network. A customer may decide to pass the username and password to friends. Friends can then access the resources without having paid for them. Authorization systems have to provide a mechanism for minimizing this risk. An authorization system should make it possible for a client to verify whether the signer of a certain check is authorized to do so. In all these cases a secure and global system of authorization is required. Clients have to be authenticated before checking whether they are authorized to access or perform a task. The first task of an authorization system is therefore to authenticate clients. So how can clients be reliably authenticated in an open network?

There are two types of authentication schemes: simple authentication and strong authentication. In

this system we use using strong authentication and clients and servers mutually authenticate each other. It is based on public key certificates. A client is required to present her public key certificate for authentication to the server. The second task of an authorization system is to check whether the authenticated client is authorized. This is described in section 4.2. What are the requirements of an authorization system in open networks?

1.2 Requirements

Authorization system in open networks must be combined with an authentication system because users must be authenticated before verifying their authorization to access resources or perform certain functions. Authentication and authorization should be done by separate systems, because the same authority does not usually create authorization and authentication information. The system must be secure so that people can trust it. It should be possible to delegate rights and privileges to other entities. It should be easy to administer which implies that an authorization system should have a user-friendly interface. The system should be scalable and efficient, because it is used in global systems where delays are not acceptable. It should support distribution of authorization elements. Authorization in open networks should be flexible supporting alternative authorization policies.

1.3 Authorization Policies

An authorization system is based on authorization policy of an organization. An authorization policy specifies rules for accessing objects or performing certain actions. This policy can be specified in terms of access control lists, capabilities, or attributes assigned to subjects, objects or both. Policies are usually described by access control models. An access control model is an abstract description of an access control system and its main goal is preventing unauthorized access to resources of a computer or information system. An access control model comprises the following items: a target, which is the object to be accessed; an initiator, which is an entity wishing to access the target and an access control function, which uses access control information to decide whether a subject can access a target. Access control function passes its decision to an access control enforcement function, which provides access to the target information or prevents it based on the output of the access control decision function.

Organization of Sections

Section two covers current approaches. Section three deals with the principles of an authorization

system in open networks. Section four describes a prototype of the authorization system. Section five briefly discusses conclusions.

2 Current Approaches

2.1 Some Solutions on Restricting Access

Authorization in open networks can be based on IP addresses and domain names [4] in which case a server examines the incoming request and grants or denies access depending on the IP number or domain name. IP-based authorization is not suitable for mobile clients and it does not accommodate dynamically allocated or shared IP addresses. This type of authorization is not secure, because today it is relatively easy to forge IP numbers. The system is vulnerable to DNS spoofing and IP spoofing where an attacker takes control of the DNS host-names' lookup system. As a result a server can be led to believe that it is talking to a trusted host. How can one verify whether an IP address is genuine? One way is to extract the IP address and then double-check with the DNS system of the client. A request can be made to the DNS to return the host name of the IP address to be checked. Then another request is made to the DNS system to return the IP number of the host name returned in the previous request. If these match then the IP address is most likely genuine.

It is also possible to minimize the problem by using firewalls, which use reliable DNS lookup. But how can one determine whether a DNS lookup is reliable? Are there any trusted and reliable DNS lookups today? Can firewalls be trusted? These systems must be properly configured in order to function correctly and not all firewall administrators are competent in this area.

Authorization can be based on certificates. When a user requests a service, she presents a digitally signed certificate together with the request. A server grants access if certificate is valid. To be valid means, that the chain of certificates has been validated and that the certificate has not been revoked.

2.2 Role-Based Access Control (RBAC) System for Securing a Web-based Workflow

Ahn, Sandhu, Kang and Park [2] describe a way to add a RBAC system to an existing web-based workflow system. A web-based workflow system consists of an interface for clients, a gateway to external services, a tool for protocols, and workflow tool for descriptions and enforcements, where

activities are performed in coordination. Different servers execute different tasks. These systems provide only low-level security services such as simple authentication. Authentication and authorization security services are based on public key certificates. The system uses HTTP protocol for client-to-server communication and uses CORBA's network addressing protocol for server-to-server communication. Different roles are attached to each task. Users' identities are verified and then checked whether authorized to perform tasks, which they request Role-Based Access Control (RBAC) model in this system has a set of roles, a set of permissions, and users. This model supports role hierarchies. Permissions are assigned to roles and users may have different roles. Users can have one or more roles. A role can be assigned one or more permissions and vice versa.

RBAC system consists of three major components: a workflow design tool, a role server and a Web-based workflow system. The workflow design tool is used for the administration of the system: generating roles, building role hierarchies, assigning roles to tasks, specifying flows of information and relationships among tasks, and for passing information to the role server. The role server has two components: a user-role assignment component and a certification server. The functions of user-role component include assigning users to roles, and creating and managing role hierarchies and databases. The certification server is responsible for verifying users' identities, fetching users' information from databases and issuing certificates with users' role information. The workflow system contains Web-based task servers. A task server approves authorization to a client based on the information found in user's certificate. The client is given authorization during the establishment of SSL session between a client and a task server. The Web server asks for a client certificate during SSL handshaking procedures. Client sends a certificate to the server. The server verifies the identity of the client. The server extracts authorization information from the client's certificate and checks whether this client is authorized.

The advantage of this system is that very little changes need to be made on the server side and no changes on the browser's side. If one of Web servers gets manipulated, it doesn't cause the system to stop, because servers are doing multiple and different tasks. The disadvantage of this system is that both authentication and authorization information are based on public key certificates. Authorization and authentication information can be set and updated by different authorities. It is also inconvenient with

respect to policy management, because different authorities can have different policies. Validity of authorization information and authentication can also be different.

2.3 One-Shot Authorization System using Smart Cards

Au, Looi and Ashley [1] present an authorization system based on smart cards. This system can be used in coordination with any existing authentication system and it can authorize clients across multiple domains. In one domain the system consists of three components: a client workstation, a security server, and an application server. The client workstation is connected to client's smart card reader. On this workstation there is a program called Authorization Token Manager. This program communicates with an application server and the administrator of the application server installs it on the client side. This program retrieves one-time authorization tokens, verifies them and stores them in the smart card together with private keys and other information. Client's smart card authenticates remote servers, verifies authorization tokens and also creates session keys. After using these one-time tokens the program renews them. Security server contains two modules: an authentication server and an authorization server. An authentication server verifies identities of clients. An authorization server performs authorization services. The security server communicates with an application server to get initial and updated authorization information. It also communicates with the workstation to exchange authentication information. The application server maintains an access control list, a valid token ID list, and access control information list.

The advantage of this system is that authorization tokens are one-time, which solves the problem of replay. The disadvantage of the system is that it creates heavy traffic, because only one-time authorization tokens are issued. Another shortcoming of this system is that it is not explained how the messages are protected while in transfer, so it is difficult to determine how secure the messages are during this process.

3 Use of Attribute Certificates for Authorization in Open Networks

3.1 Attribute Certificates

An attribute certificate (AC) is a signed object containing authorization attributes of a subject. Attribute Authorities (AA) are the components responsible for issuing attribute certificates. The serial number of the client's PKI certificate, which is used for authentication purposes, is inserted in a field called *holder*. Fields of an attribute certificate according to [6] are:

- Attribute certificate information
- Signature algorithm identifier, which is an algorithm used to sign the AC
- Signature value, which is a signature of the issuing AA

The fields in the attribute certification information include:

Version: This field contains the version of the attribute certificate (AC).

Holder: This field contains the identity of the holder of the certificate, that is the entity to whom the attributes apply. It has the serial number of the owner's public key certificate, general names of the AC's owner, digest information, which can include public key, public key certificate, digest algorithm and so on.

Issuer: It contains the identity of the issuer of the attribute certificate.

Signature: This contains the algorithm that was used for signing the attribute certificate.

SerialNumber: It has a serial number of the attribute certificate.

AttrCertValidityPeriod: This field contains the validity period of the attribute certificate in the form of two dates defining a time interval.

Attributes: This field contains the actual attributes and is specified by the issuer of the attribute certificate. These attributes include service authentication information, access identity, charging identity, group, role, clearance and etc.

IssueUniqueID: This field contains additional information to help locate the issuer.

Extensions: Extensions contain some additional information about the attribute certificate: *audit identity* for audit trails; *attribute certificate targeting*, which is used to specify the number of targeted servers or services; *authority key identifier*, which is used to assist in verifications of the signature of the attribute certificate; *authority information access*, which is used for checking revocation status of a certificate; *CRL distribution points*, etc.

Attribute certificates are stored in the same way as public key certificates: in global repositories or in directory systems. Attribute certificates can be revoked. But in cases when their lifetimes are too short, revocation may not be necessary. Revoked attribute certificates can be stored in attribute certificates revocation lists. This is a list of AC's serial numbers. It must be possible to verify the authority of the issuing AA, i.e. there is a valid chain of public key certificates containing the extensions asserting AA's authority. In inter-domain environments there should be a way of translating attributes issued by other domains into the domains responsible for validating the ACs. Attribute certificates should keep all or some of its attributes confidential if so desired by clients. Attribute certificates are useful in supporting delegation.

3.2 Authentication of Clients and Assignment of Roles

When a client connects to an authorization server for the first time she is authenticated by presenting her public key certificate. This certificate is verified by validating certification chain from the authority, which issued the certificate to the top certification authority in the hierarchy. A check is also made to verify that the certificate in question is not revoked. If the certificate is found to be valid then an attribute certificate is issued to the client. Roles and clearances are given to the client and they are written in the client's attribute certificate. These roles and clearances specify authorization of the client and these specifications are stored in the policy file of the attributes authority. A reference to the client's public key certificate is also included in the attribute certificate in the field called *holder*. In this attribute there is a sub field called *baseCertificateID* and this sub-field holds the serial number of the client's public key certificate. After populating all the fields of the client's attribute certificate, the certificate is signed by the issuing AA. If the client desires to protect some fields of the attribute certificate that can be done using a secret key. The attribute certificate is then stored in the X.509 Directory or in a global database. A copy of this attribute certificate is sent to the client.

3.3 Synchronization of Roles and Authorization Attributes

When a client makes a request to access resources of a secure Web server, she presents her public key certificate. This certificate is validated as described in section 3.2. If validation is successful then the serial

number of this public key certificate is used to pull the client's attribute certificate from the directory or from the global database. If client's AC is not found at the server or if the database or X.500 directory is down, then the client is requested to send her AC to the Web server. Every resource in the secure Web server has a security label. Labels are attached to resources by using S/MIME. S/MIME is a standard for encapsulating MIME documents and provides services like confidentiality, integrity and authentication. Confidentiality is a security service, which protects resources from illegal read, illegal access, deletion, sabotage and so on. Integrity is a security service that protects resources from illegal modification, deletion and etc. The resources are stored in the security Web server in encapsulated forms. The security labels that are attached to resources specify in the policy file which roles and clearances can access the corresponding resources. The security label has a list of all roles, which can be granted access. The policy file contains information on security classifications and categories. It can contain information mappings among different security policies. If a policy of a company changes then it is enough to update the policy file without needing to change other modules. A Policy Creation Authority (PCA), which is a trusted entity, signs the policy file. Security labels and clearances have policy identities, which are references to the policy files in which they are specified. The policy file contains lists of security classifications and categories and all allowed combinations of them. All messages between a client and a secure Web server are protected using S/MIME, SSL or other secure protocols.

3.4 Enforcement in the Authorization System

Decisions to grant access to the secure Web server's resources are based on the policy of the AA. This policy is created by the Top or Root certification authority and all the entities under this root certification authority use this policy. Roles, clearances, ranks, security labels and other attributes and information are specified in this policy file. The attribute certificate of the client is pulled from the global database or from the X.509 Directory. The security Web server must verify the attribute certificate by verifying the signature of the attribute certificate. The validity of the AC must also be checked. The subject in the attribute certificate, the AC's issuer, and the complete certification chain is validated. A local certification authority, as explained in section 3.5, certifies the AA. The client's AC contains clearances or roles of the client. These attributes specify the authority of the client. Access

control decision function takes as parameters, a policy file, a security label, and an authorization set and this set includes a clearance, a role and other parameters. Access is granted if the client's attribute certificate is verified and if the client has a clearance or a role that matches the security label of the requested resources.

3.5 Management Infrastructure

The system uses the X.500 authentication framework. This system uses certificates-based authentication. Clients are required to have public key certificates before being authorized to access or perform actions in the authorization system. Certification authorities (CA) certify attributes authorities (AA), which issue attribute certificates. The complete system is shown in Figure 1. At the top there is a trusted root certification authority. Below this root CA there is one or more intermediary certification authorities depending on the complexity or size of the organization. The last CA in the hierarchy is a local CA. This is responsible for certifying the Attribute authority, managing public key certificates to clients, managing keys, revoking certificates and so on. At the root CA there is a Security Policy Information File (SPIF) for the entire system. This file contains the policy for the whole system. Every certification authority has a certificate, which contains an extension called cAClearanceConstraints. This extension enables authority to act as an Attribute Authority (AA). The root CA issues a self signed certificate to itself. It then issues certificates to the lower entities. If the root CA belongs to a company then this company can have middle certification authorities in different countries where it has offices or its business.

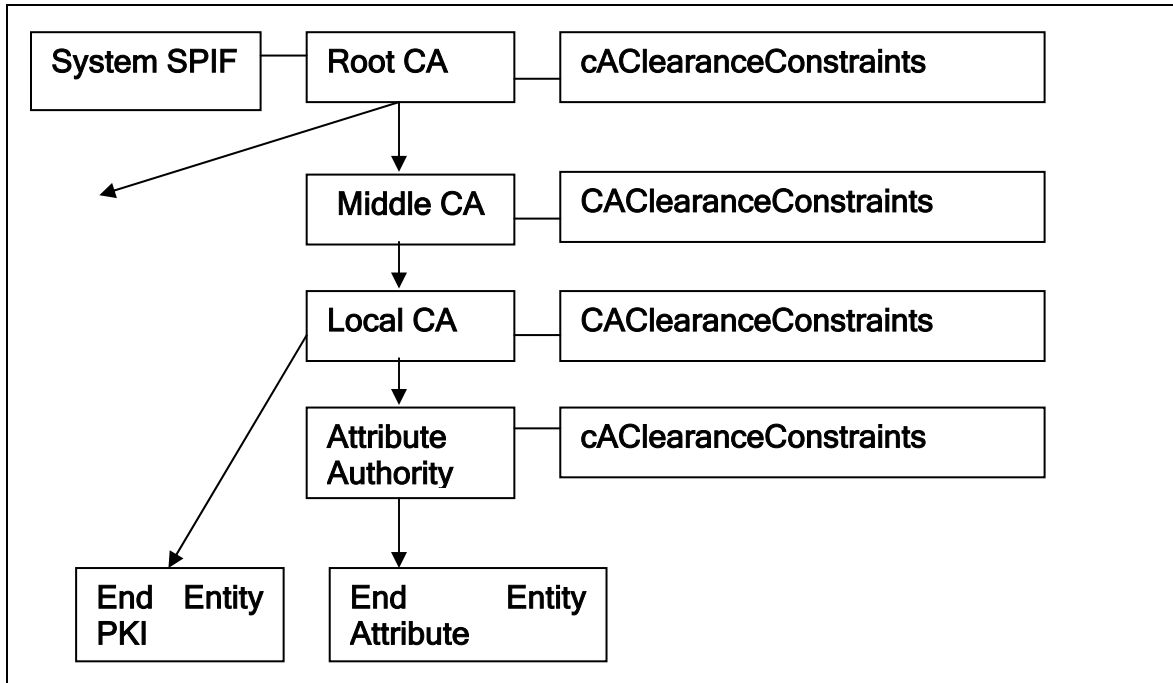


Figure 1: Model of System Components

Certificates issued to lower entities have to be verified by checking the signatures of certificates. The whole chain up to the root CA has to be validated. The local CA issues a certificate to the AA, which in turn issues an attribute certificate to the end entity. The policy file, SPIF, has to be signed by the root CA so end entities must verify the signature before using it. In cases where there are different root certification authorities and belong to different organizations then root certification authorities are required to cross certify each other so they have to issue certificates to each other and these certificates will contain the corresponding cAClearanceConstraints extensions.

3.6 Delegation of Attributes

Delegation of attributes is done with the help of a file called *attribute* in the attribute certificate and also with the help of an extension in the AC that is called *authority information access*. Authority information access has an IP address of the directory where the issuer of the attribute certificate may be found. This extension can also store an IP address of the directory that has the AC of the upper entity that delegated attributes to the lower entity. When the Web server receives a request from a user it can authenticate her as described in section 3.3 and if authentication of the user is successful, the Web server will retrieve the user's AC and check its

validity as discussed in section 3.4. If attributes are delegated then the attribute's value will be *delegated set of attributes*. The Web server will thereafter get the AC of the delegating entity from the directory whose IP address was in the authority information extension. The AC of the upper entity will be verified as discussed in section 3.4. The user will be authorized if the AC of the delegating entity is valid.

4 Implementation of a Prototype

This prototype is based on geotronics [7] library suite. The RBAC is implemented using access control library in the following way. It is specified in the policy file, SPIF file, as described in section 3.4, so that all the roles are given security categories. Categories are authorities to perform different functions or access different objects in the secure Web server. Every security label has a list of roles, which are authorized to access certain resources or perform the desired actions. Every clearance in the attribute certificates contains a list of the roles, which can be granted access.

4.1 The Access Control Library Suite

This authorization system uses the access control library [3] and it consists the following libraries.

SNACC. This is a high performance ASN.1 to C/C++ Compiler. This library contains an ASN.1 compiler for encoding and decoding data structures.

S/MIME Freeware Library (SFL). This library provides support for cryptographic functions like signing, verifying signatures, protecting messages and so on.

Certificate Management Library (CML). This is used to verify the certification paths.

The Storage and Retrieval Library (SFL). This library is used for maintaining the database for certificates. SFL is used for providing functions for parsing, generating, protecting and verifying SMIME messages.

Access Control Library (ACL). This library takes care of access control decisions basing on S/MIME security labels, X.509 certificates and attribute certificates.

4.2 Implementation

There are three components in the prototype: An administration tool, a SPIF generator and a certification manager. An administration tool, AdminTool (figure 4), is used for managing roles and

S/MIME documents. The SPIF generator is used for generating policy files, SPIF.

The administrator chooses an item to be generated

from the interface in Figure 2. The administrator can choose to generate an SMIME document, an attribute certificate or a new Security Policy Information File (SPIF), see Figure 2:

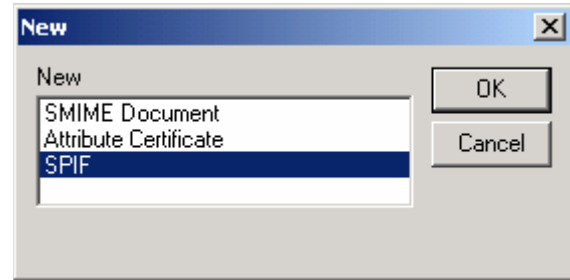


Figure 2: Choosing an Item to Generate

The administrator of the system creates a policy file as explained in section 3.4, enforcement in the Authorization System. She/he does this by activating the SPIF generator and a panel shown in Figure 3 will be displayed. In the SPIF generator, one has to specify the policy ID and a version of this policy file.

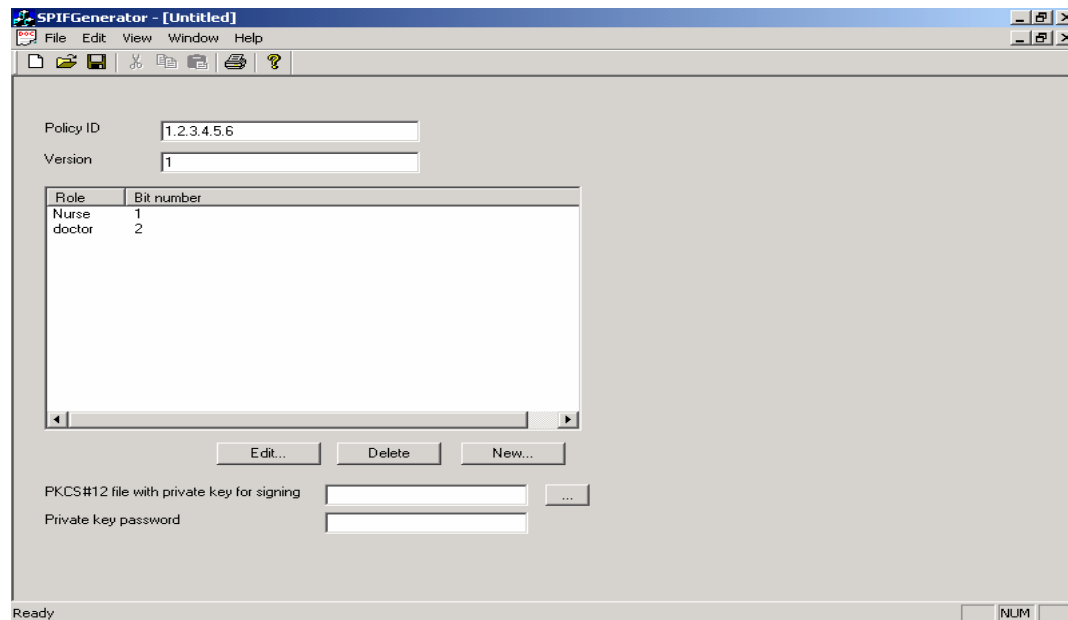


Figure 3: SPIF Generator

Then roles have to be created. After creating the roles, SPIF file must be signed using the private key belonging to the issuer of the policy file.

To issue an attribute certificate as discussed in section 3.2, recognition of clients and assignment of roles, an administrator selects option attribute

certificate from the interface in Figure 2. Thereafter the administrator selects the policy file and public key certificate for authentication purposes as described in section 3.2. Different fields like serial number, validity, roles, etc, are populated in the attribute certificate. The attribute certificate is then signed. Before storing the attribute certificate to the database, trusted certificates must be added to the database or to the directory system. These certificates

are necessary for certificates chain validation as discussed in section 3.2. This is done using Certificate Manager interface, shown in Figure 4. Attribute certificate can then be added to the database or to the X.500 Directory.

The next step for security administrator is to attach security labels to resources (in this case Web documents) as described in section 3.3. To add documents to the Web server, the administrator selects *SMIME* Document option from the *AdminTool* panel and this panel shown in Figure 5.

Then he/she selects a document to be encapsulated and the corresponding SPIF to be used. The administrator then specifies the roles, which can access this document.

The private key for signing the security label must be specified. When a client requests to access a site on the Web server, the server expects client's public key certificate. The Secure Socket Layer [8] is used for establishing secure sessions between the client and Web server. SSL is a system for securing messages while in transfer. The server checks whether client's public key validates the client's

digital signature as discussed in section 3.2. It also checks whether today's date is within the certificates validity period. It also checks whether the CA that issued client's certificate is a trusted CA and also whether the public key of client's certificate issuer validates the issuer's digital signature. The server checks whether this certificate corresponds to the serial number in the attribute certificate. Then the Web server checks with the ACL to decide whether an incoming request is authorized to access the site. Web server loads the Publish dynamic library and passes the name of the selected document as a parameter to the access method of the extension. Documents are stored on the server in S/MIME format and contain security labels as described in section 3.3. The extension function fetches user's attribute certificate from the X500 Directory and compares the role in it with the security label of the requested document. If the client is assigned the roles contained in the security label of the document, the document will be transferred to the client. If the client is not authorized to access the file, he/she will get an http "404 Not found" response.

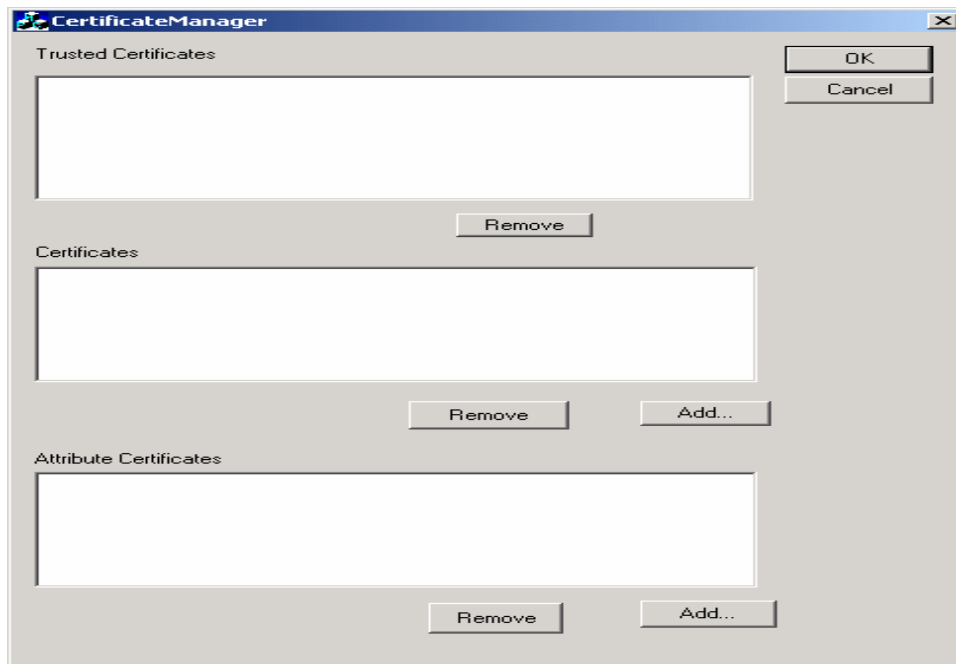


Figure 4: Certificate Manager

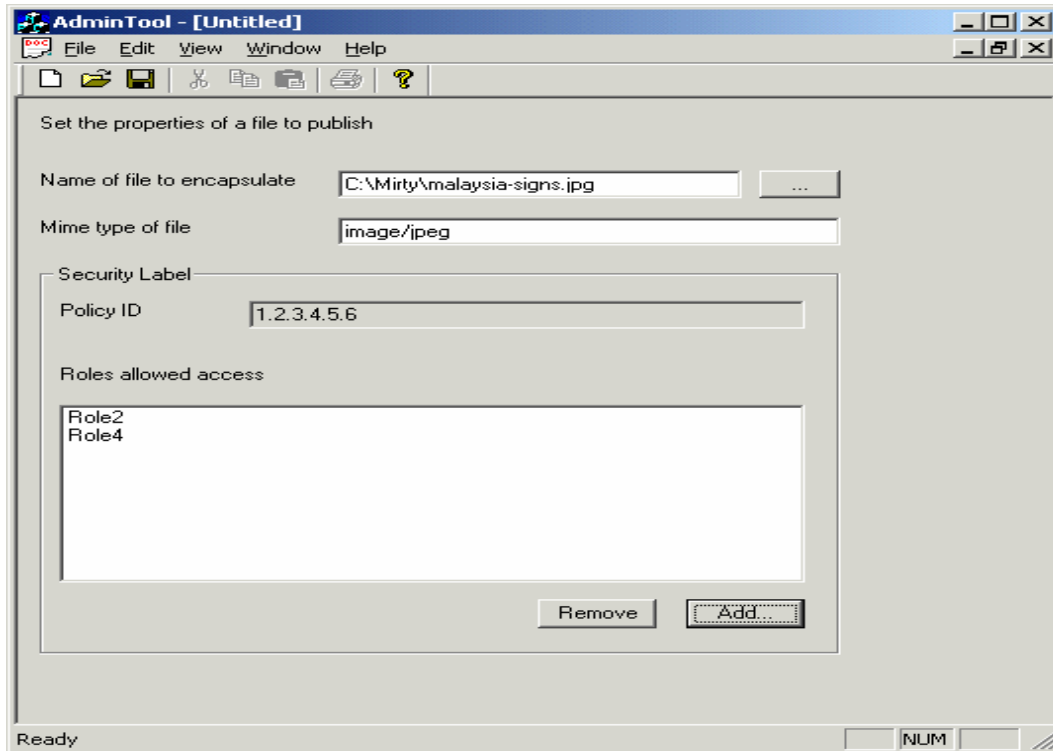


Figure 5: AdminTool

5 Conclusions

This system is flexible and interacts with other systems like PKI certification system, X500 directory system and smart card systems. Attribute certificates support delegation through an ordered sequence of attribute certificates with references to certification authorities. Attribute certificates can be used for non-repudiation services making it possible to extend authorization systems to support this service. The system separates authentication security service from authorization making it possible for authentication and authorization decisions to be made by different authorities when necessary.

References

- [1] Au, R., Looi, M., Ashley, P. Cross-Domain one-shot authorization using smart cards, *The Journal of ACM*, 2000
- [2] G Ahn, R Sandhu, M Kang and J Park. *Injecting a RBAC to Secure a Web-based workflow system*, *The Journal of ACM*, 2000.
- [3] *Access control library*, http://www.getronicsgov.com/hot/acl_home.htm, April 2001.
- [4] Oppliger, Rolf. *Security Technologies for the World Wide Web*, 2000.
- [5] RFC 2222, SASL, www.ietf.org/rfc/rfc2222.txt

- [6] Farrell, S., Housley, R. An Internet Attribute Certificate Profile for Authorization, <http://www.watersprings.org/pub/id/draft-ietf-pkix-ac509prof-07.txt> (work in progress), June 2001.
- [7] Getronics Government Solutions, <http://www.getronicsgov.com/>, January 2002
- [8] The Secure Socket Layer, <http://home.netscape.com/security/techbriefs/ssl.html>, January 2002.



Web Based Generic File Protection System

Bimali Arsakularatne, Kasun De Zoysa, Rasika Dayarathna

Department of Communication and Media Technologies,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7, Sri Lanka.

E-mail: wbimsdfa@yahoo.co.uk, kasun@cmb.ac.lk, rasika@cmb.ac.lk

Abstract

This document describes a web based generic file protection system, which uses simple cryptographic mechanisms to provide much more stronger security services.

Web technologies are used instead of a standalone approach since it does not require installing any additional software. In addition, a Web based approach allows the users to store critical information in a secure server instead of the local hard-disk. This not only protects critical data, but also enables the user to access these data from anywhere in the world.

At present, the following generic security services are provided for registered users:

- **Integrity Checking Service:** used to detect unauthorized alteration of file systems.
- **Digital Signature Service:** useful for providing authorization and authenticity of documents.
- **Encryption Service:** provides protection against unauthorized access.

Keywords: Digital Signatures, Encryption, Data Integrity, Document Protection

1. Introduction and Motivation

There are various types of software developed to ensure the security of a computer system. Among these are virus scanners, intrusion detection systems, firewalls etc. Even though a computer is secured using all or some of these software, can we make sure that it is adequately protected? For example, how do we know that hacker did not enter the system and altered the file-system without leaving any evidence?

Data in your computers can be compromised in the following ways.

- **Eavesdropping:** Information remains intact, but its privacy is being compromised. For

example, someone could secretly read your personal emails without altering them.

- **Tampering:** Information is changed or replaced without your knowledge.
- **Impersonation:** Information passes to a person who poses as the intended recipient.

There are three cryptographic mechanisms, if used effectively, would ensure data protection in both storage and transmission of the data. They are described below.

Integrity Checking

The file system of a computer contains all of the long-lived data in the system including all user data, application data, system executables and databases. Therefore the file system is one of the usual targets of an attack. Motives for altering system files are many. Intruders could modify system databases and programs to allow future entry. System logs could be removed to cover their tracks or discourage future detection. Compromised security can lead to faulty services. Therefore the integrity of a file system should be closely monitored.

Integrity checking is most useful when tampering is so carefully and ingeniously carried out such that common means such as the 'last modified date' of a file or logs cannot identify any changes.

Digital Signatures

Although the use of digital signatures is not very popular in Sri Lanka and there is no legal enforcement based on digital signatures yet, we should anticipate and be prepared for the future. The value of digital signatures is immeasurable when it comes to e-business whether or not the business is web based. For example, from their experience of dealing with organizations of all sizes, it became apparent to DocumentFlow (a major software company in the

United States) that designers wanted to issue AutoCAD drawings to clients and contractors, but were reluctant to do so for fear of modifications to those drawings after approval [4].

Data Encryption

Data encryption plays a major role in e-business as well as within an organization in the process of passing messages between users. We often take email for granted and send confidential messages via email. But if we were to carry out the same process manually, we would take so many security precautions.

2. Background Theory

We start by introducing some concepts that are used throughout this paper.

2.1 Security Services

Data Integrity

A level of assurance that ensures information has not been deliberately or inadvertently modified or replaced in transit or storage.

Authentication

The process of proving one's identity.

Non-repudiation

The capability to demonstrate that an action such as the sending of a message was performed by a person with a particular identity.

The actual process of how these tasks are accomplished will be discussed consequently.

2.2 Secret Key Cryptography

In some encryption algorithms, the encryption key and the decryption key is the same, or the decryption key can be calculated from the encryption key. These algorithms are known as secret key algorithms (or private key algorithms/symmetric key algorithms). The encryption key must be kept secret and the sender and receiver must coordinate the use of their keys.

Following are some popular secret key algorithms [1]:

DES

This is the Data Encryption Standard algorithm. There are known ways to attack this encryption. But they require a lot of computing power to do so.

DESede

This is also known as Triple-DES or multiple-DES. This algorithm uses multiple DES keys to perform three rounds of DES encryption or decryption. The added complexity increases the time required to break the encryption as well as the time required to encrypt or decrypt data.

PBEwithMD5andDES

This algorithm uses a password, a byte array known as a salt and an iteration count along with an MD5 message digest to produce a DES secret key. This key is used to perform DES encryption and decryption.

Blowfish

This algorithm is best used in applications where the key does not change often. It requires a lot of memory.

2.3 Public Key Cryptography

One of the most important breakthroughs in cryptography during the 20th century was the development of public key encryption. Public key algorithms (or asymmetric key algorithms) use separate keys for encryption and decryption. The encryption key is referred as the public key and the decryption key is called the private key. That is because when data is encrypted using the encryption key, you need the decryption key for the decryption of the data. Hence one can make his/her encryption key known to the world at large and therefore it is called the public key. On the other hand, one should protect his/her private key because data can be decrypted with it. The important feature in public key cryptography is that the sender and receiver do not have to share keys.

Figure 1 shows how public key cryptography works.

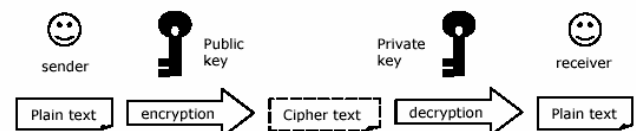


Figure 1: Operation of public key cryptography

RSA (Rivest, Shamir, Adleman)

The most famous public key algorithm – developed in 1977. The following is a summary of how RSA works [2]:

- Two large (100 digits or more) prime numbers p and q are generated with $n = pq$
- A public key e is selected as an integer such that e is relatively prime to $(p-1)(q-1)$
- The private key d is computed such that $ed \bmod ((p-1)(q-1))$ is 1
- Encryption is performed on plaintext numbers m that are smaller than n by calculating $m^e \bmod n$
- Decryption is performed on cipher-text c by calculating $c^d \bmod n$

Disadvantages of RSA

- RSA is 100 to 1000 times slower compared to secret key algorithms [2]
- Although RSA has not been broken, if an efficient way to factor large numbers was discovered, it could be easily broken.

2.4 Message Digests

Message digests are used to secure data integrity. In other words, to detect whether data has been modified or replaced.

A message digest is a special kind of function referred to as a one-way (hash) function. A one-way function is easy to calculate, but difficult to reverse. Message digests take messages or data as inputs and compute values referred to as hash values that are used as fingerprints to the messages.

Good message digests have the following properties.

- Given a particular hash value, it is computationally infeasible to compute the message that produced that value
- It is computationally infeasible to find two messages that yield the same hash value.

Examples of Message Digest Algorithms

- SHA1
- MD5

2.5 Secure Message Digests

A secure message digest is called a Message Authentication Code (MAC). A MAC has the property that it cannot be created solely from the input data. It requires a secret key that is shared by the sender and receiver. Hence an intermediate party cannot change both the data and the MAC without the receiver detecting that the data has been corrupted.

2.6 Digital Signatures

Digital signatures are mainly used to prove to you that a message sent to you is created by a particular individual or an organization. If the receiver can verify the digitally signed message from the sender, then the receiver can make sure that the contents of the message are correct and authentic.

Digital signatures have the following properties similar to real world signatures [2].

- Unforgeability – because the signer uses his private key to sign, only he can sign with that key
- Verifiability – because the signer's public key is openly available, anyone with access to the message and signature can verify that the message was signed by the signer and that neither the message nor the signature has been altered.
- Single use – A signature is unique to a particular message.
- Non repudiation – After a signer has signed a message and the message and signature have been sent to others, the signer cannot claim that he did not sign the message.
- Sealing – A signed message is digitally sealed. It cannot be altered without invalidating the signature.

2.7 Digital Certificates

When you use public keys of others to encrypt data or verify their digitally signed documents, there is no way to make sure that the public key belongs to the particular person that you are referring to. This is where digital certificates come in. Digital certificates are messages signed by a Certification Authority (an entity that is trusted to verify that other entities are who they claim to be) that certify the value of an entity's public key.

Figure 2 illustrates the use of digital signatures.

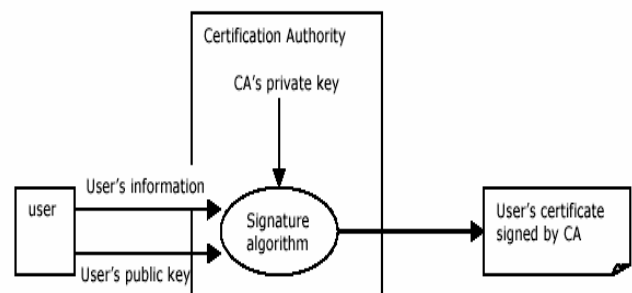


Figure 2: Use of digital signatures

2.8 Signed Applets

By signing an applet an organization or an individual can indicate that the organization has reviewed the applet for security and believes that the signed applet is free from security defects. The signature also implies that the organization/individual takes responsibility for the applet in cases where there is a security malfunction. Signing also provides the user with a mechanism for verifying that a signed applet originates from a particular organization and has been delivered to the user without modification. For these reasons, a user can determine that he or she is able to extend a certain level of trust to an applet that is signed by a reputable organization.

An applet is signed in two steps. In the first step, the applet's class files are archived in a JAR file. In the second step the JAR file is signed with a signing tool.

3. Literature Review

3.1 Review on Integrity Checking

Integrity analysis focuses on whether some aspect of a file or object has been altered. This often includes file and directory attributes, content and data streams. Integrity analysis often utilises strong cryptography mechanisms called **message digest** (or hash) algorithms that can recognize even subtle changes.

In simplest terms, a database is created with some unique identifier for each file to be monitored. By recreating that identifier (which could be a copy of the entire file contents) and comparing it against the saved version, it is possible to determine if a file has been altered or not. Furthermore, by comparing the entries in the database it is possible to determine if files have been added to or deleted from the system.

The file contents themselves are not usually saved, as this would require too much disk space. Instead, a checklist can be used. The checklist would contain a set of values generated from the original file (usually including the length, the time of last modification, owner etc.). It should be periodically regenerated and compared with the saved version and note what the differences are. However, changes may be made to the contents of files without any of these values (the checklist values) changing from the stored values [3]. Therefore to efficiently detect changes in the file system, values should be calculated from the contents of the files itself. If this value depends on the entire contents of the file and is difficult to match for an arbitrary change to the file, then storing this value is sufficient. This fingerprint or signature of the file can be saved instead of the file itself.

- The signature files used should be computationally simple to perform but infeasible to reverse.
- It should signal if the file changes
- It should be sufficiently large as to make chance collision unlikely.

3.2 Issues faced by an Integrity Checking Software

Database Issues

If the database can be updated after every authorized file addition, update or deletion, this prevents the change being reported in future checking events and hence reduces the complexity of the report issued.

But the database should also be secured from tampering. If a copy of the database can be kept remotely or on a removable disk, this problem can be solved.

Signature spoofing

Intruders could modify a file and remain undetected in an integrity checking scheme using file signatures if the file can be further modified to generate the same signature as the original. Two methods for finding such a modification are:

- Brute force search
- Inverting the signature function and spoofing the signature function

For these reasons, message digest algorithms become valuable as integrity checking tools. Message digests are usually large, often at least 128 bits in length and computationally infeasible to reverse and carry out a search.

Duplicate Search

The large number of collisions for the 16 bit signatures and the absence of any collisions for the 128 bit signatures confirm the expected observation that larger signatures are less likely to collide by accident.

Other Issues

The Integrity checker should only report, and not effect changes. Although a user could use the tool's output to drive changes, the tool itself would not provide any explicit means of making alterations to the system.

3.3 Review on Digital Signatures

There are many types of software developed so far to support the use of digital signatures. But a major flaw in them is that they do not support signing all types of

files. Some software supports digital signing of specific file types while others supports signing of document types only.

For example, *Adobe Digital Signature Architecture* provides a plug-in to sign PDF files [5]. *CADSign* enables AutoCAD file signing [4]. *Approve Desktop* claim to sign any type of document such as Word, Excel, PDF, AutoCAD, XML, HTML etc.

A reliable digital signature scheme should be able to sign any type of file whether it is a document type or an executable although it is true that documents are the most frequently used file type that needs digital signing when businesses and organizations are concerned.

Since a file can be signed by many signers, we should be able get information about each of the signatures on a file. This information should include the identification of the signer, signature date and time etc.

3.4 Review on Data Encryption

There are many types of software that support data encryption. But since it is one of the most essential components of a good file protection system, encryption facilities will also be included in the system being developed.

Some encryption algorithms can be broken in a matter of hours; some would take many years. Others would take several times as the anticipated lifetime of the universe to break, given machines many times more powerful than the ones in use today. Of course, the price you pay for more security is the encryption time, among other things. If the data will be useless in an hour, you do not need an algorithm to protect it for your lifetime.

Some algorithms are prohibitively slow for common use. If you need a Cray mainframe to encrypt and decrypt the data in a reasonable time, it probably is not a good choice for an applet.

4. The System Architecture

The system is web-based so that everybody can enjoy the security services provided. A web site is developed which supplies the following services to its users.

- Obtaining membership with the server
- Generating symmetric and asymmetric keys
- Digitally signing documents and uploading them to the server
- Verifying digitally signed documents
- Providing encryption and decryption facilities
- Scanning for integrity of specified files
- Updating the integrity status of selected files

- Calculate hash value of specified files

4.1 Why web based?

By allowing the users to download the programs as signed applets, it can be guaranteed that they have access to the latest version of the software and that it is not altered by some third party. Another important advantage of a web based system is that users can access the system from anywhere in the world to use the provided services.

4.2 Remote Database Storage

Users may find their computers as an insecure place to store critical data items such as keys, digital certificates etc. Our system provides a facility to store their critical data items at a remote database if they prefer to do so. The remote database can be located at the same machine as the web server or at some other location.

It should be pointed out that a remote database is vital for the integrity checking operation. The reason for this is that if the message digests were stored locally, a hacker who changes a file can change the corresponding message digest so that a change would not be reflected. Therefore the message digests should be stored more securely.

4.3 Client Side

Applets are used in the client end of the system. A bulk of the processing is being done in the applets. This helps to reduce processing weight on the server as well as unnecessary network traffic since most operations do not need to access the server. For example, encrypting a file can be processed totally at the client side if the necessary keys are located at the client. Otherwise, if the keys are located at the server, the applet just has to download the necessary key from the server and carry out the operation.

Since ordinary applets restrict many vital operations (such as file accessing at the client side) signed applets are used. All the operations that can be performed by a normal java application can also be performed using signed applets.

4.4 Server Side

The server side will be manipulated via Java Servlets. Since Servlets provide multi-threaded access by default this was a natural choice. The servlets have to handle database access depending on the call from the clients.

4.5 Data Transfer

The applets generate a symmetric key at initialization. Then this key is encrypted with the server's public key and is sent to the server. Thereafter the applet and the server communicate with each other using the symmetric key to encrypt data (see Figure 3).

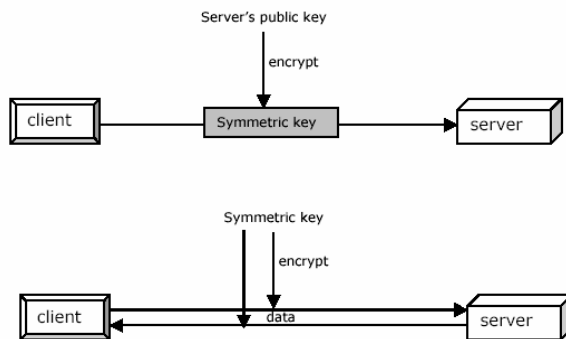


Figure 3: Data transmission protocol

4.6 Integrity Checking Process

- The integrity checking is handled using message digests.
- Initially, the user has to store the state of his file-system when the file-system is at a safe state (i.e. the user is certain that the file system has not been changed without his/her knowledge). At this stage, a message digest for each file is calculated and stored at the server database (see Figure 4).

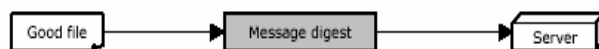


Figure 4: Update of file status

- When the file system is to be scanned for integrity checking, a set of message digests will be calculated for each file once more. Then these message digests will be compared with the ones stored at the remote server. If these two are different for a certain file, then the corresponding file has been changed (See figure 5).

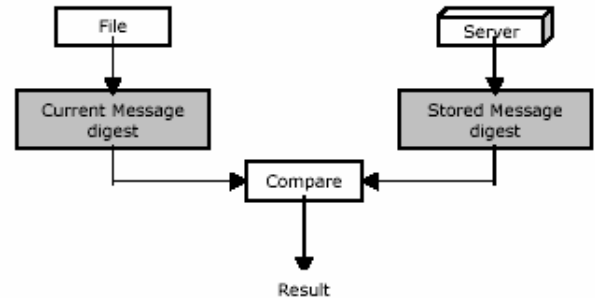


Figure 5: Integrity checking

In addition to the integrity checking process, the system also supports document signing and verification as well as documents encryption and decryption.

5. Functionality of the System

5.1 Obtaining Membership with the Server

Every user of the system has to register with the system and obtain a unique username. To register, users must provide personal information such as full name, address, nationality, occupation etc. Users must provide a password which will be used in subsequent login sessions.

5.2 Generating Symmetric and Asymmetric Keys

User selects the type of key that should be created. E.g. Symmetric/asymmetric. Next, the key algorithm to be used and the location where the key should be stored has to be given. Finally, the keys are generated. Key name is validated and checked for duplications.

5.3 Signing and Verifying Documents

User selects the file to be signed and the private key to sign with. This private key can be one which is stored locally or the user can download one of his/her pre-generated keys from the server. After that the user has to choose to either save the signed file or to upload it to the server. The system will act accordingly.

In the case of verification, user selects the signed file and the relevant public key or certificate to verify with. After the verification, the system indicates to the user whether the verification was successful or not.

5.4 Encryption and Decryption Facilities

User selects the file to encrypt. Then chooses the key type (symmetric or asymmetric) and the key algorithm

(DES, DESede etc. or RSA). Then the applet will save the encrypted file as specified by the user.

In decryption, user selects the encrypted file and the relevant key. The system will decrypt the file and save it on the hard disk.

5.5 Scanning for Integrity of Specified Files

User selects the files that should be scanned. He is given 4 options to:

- Scan the complete file system
- Scan critical files
- Scan selected partitions
- Scan selected files and directories

The results will be displayed in a report. User can update the status of the scanned files by clicking a button on the report.

6. Featured Attributes of the System

The special features of this system that make it outstanding from the rest are discussed below and the above-mentioned qualities will be addressed as appropriate.

6.1 Remote Storage of Critical Data

Many computer users have to share the same computer with several people – may it be at the office, college or at home. In such situations, they face the problem of not being able to protect their important files. Some machines, even if used by one person, if it is on a network, could be open to trespassing of others. Suppose they choose to hide their data using some kind of key – based encryption method, still their keys would be in danger. Likewise, if the hash values that represent the file system integrity are stored in such an unsafe computer, there would be no use of integrity checking either.

Therefore it can be seen that it would be very convenient if there was a safe place that we can store our critical data items and retrieve them whenever we want to. The developed system provides this facility of remote storage of critical data. Of course, the system administrator has to ensure that the machine with the database is fully protected and the users have to be confident that their data are safe at the server.

6.2 Platform Independence

Since this system is developed using Java applets, it is platform independent and can be executed in any operating system including Unix, Linux, Windows platforms etc. This is a very important issue because most networks contain nodes with different operating

systems and if the security services cannot be utilized in one of them, then the whole network could be compromised.

6.3 Use of Key - Based Encryption

Many software that provide data encryption facilities operate with the use of passwords. But there are freely available software that can be used to crack a particular password. Key based algorithms are more powerful because it is not possible to attack key based algorithms easily. Even though there are known ways of attacks for key types such as DES, such attacks need very powerful computers which are not available for the ordinary user.

6.4 Installation is not necessary

No software installation is necessary for the operation of the system. This is very convenient in contrast with the software that have to be installed in each and every computer. Another point to note is that by the use of signed applets, the software itself is protected from unauthorized alterations and that the users can ensure this by the use of digital signature verification.

6.5 Can sign or encrypt any type of file

Most currently available software can only sign several types of files, specially document types. (See section 3.2). But the developed system can sign or encrypt any type of file.

6.6 Ease of Integration

There are no complexities in system integration. The server contains the servlets that can be executed using any servlet enabled web server. After the initial installation of the server, all the other users can access the system through the web via their usual web browser. The important thing to note is that the JDK 1.4 plug-in is a necessity for this system to function properly as the earlier versions of Java do not contain the cryptographic extensions.

6.7 User-friendliness

The system is user-friendly and is usable by the average person. The good report structure provided in Integrity scanning is a good example for this and it is comprehensible by the average person. Tool-tips and help messages are given at necessary points and an online user manual is also available.

In general the system provides many facilities together with the ones which were mentioned in this

section and users just have to log on to the website to exploit them.

7. Conclusion

At present, Computer file protection is a very important requirement in all sorts of computer systems. The Web Based Generic File Protection System provides security services required for protection of computer files. The developed system is only a first attempt and it could be improved to a professional standard together with the enhancements discussed in section 8. Once improved, it can be used either in the Internet or within an intranet by an organization to protect their data and to carry out their day-to-day functions more confidently.

8. Future Enhancements

At the current stage, the system provides many services, but it can be enhanced making further improvements as discussed below.

- As stated in section 2.7, digital certificates provide a way of ensuring that a particular public key belongs to a specified party. The file protection system can be improved to support signature verification by digital certificates as well as public keys.
- Currently the system can scan only the machine by which the applet is being accessed. When it comes to scanning a large network, scanning machines one by one becomes a tedious task. The system can be improved with the use of mobile agent technology, such that a complete network could be scanned with one attempt.
- The system would be more useful for system administrators if there is a facility where one can enter the IP address or the location of a machine located remotely and then supervise the scanning remotely.
- Only few encryption algorithms are implemented in the system. It can be extended to support more.

References

1. "Java Security", 2nd edition, Scott Oaks, O'Reilly & Associates, Inc. May 2001
2. "Java Security Handbook ", Jamie Jaworski, Venkata S. Chaganti, Paul J. Perrone, Jamie, Macmillan, USA, September 2000
3. Design and Implementation of Tripwire: A file system integrity checker – Gene H. Kim and H. Spafford, Department of Computer Sciences, Perdue University

4. "Signed Off", Article from CAD User Magazine (www.cadserver.co.uk)
5. "Acrobat Digital Signature Overview", Technical Note #5400 (Adobe Systems Incorporated www.adobe.com)



Framework for modelling of tacit knowledge -Case study using Ayurvedic domain

D.S. Kalana Mendis
Department of Information Technology,
Advanced Technical Institute,
Labuduwa.
kalanaatil@mail.com

Asoka S. Karunananda
Department of Mathematics & Computer Science,
Open University of Sri Lanka,
Nawala.

U. Samarathunga
Institute of Indigenous Medicine,
University of Colombo,
Rajagiriya.

Abstract

A research has been conducted develop a framework for tacit knowledge modelling, which is of great interest today. Here, we have considered domain of "Ayurvedic" medicine as a case study for domain with tacit knowledge. A questionnaire used to classify individuals in Ayurvedic has been studied and found that the classification is still vague, subjective and cannot be addressed using traditional technique like Principal Component (PC) analysis We have developed an approach to model such tacit knowledge using PC and Fuzzy Logic that has been linked with Expert system technology. PCA is the standard scientific approach finds any dependencies in a data set. Therefore first, we have used that technique.

This research work has produced a tacit knowledge-modelling framework, which is delivered as an added feature for an ordinary expert system shell.

direct access to resources in this field. In our research we have considered the sub area of Ayurveda called classification of individuals, which a key basis for deciding on treatments for patients according to Ayurvedic medicine.

In general, individual differences have not been taken into consideration for prescribing drugs for diseases. However, there is evidence that different drugs have different effects depending on individuals. For example, certain drugs become allergic to some people. Nowadays, there is a growing interest in study of prescribing treatments according to individuals. These works have been based on DNA technology for identification of individual characteristics. At present, DNA technology is expensive and the technology is still developing. Further this knowledge is not readily available for the ordinary use. On the other hand knowledge about DNA is very much supported by modern scientific evidence and cannot be treated as an example for tacit knowledge.

1. Introduction

All knowledge is either tacit or rooted in tacit knowledge (Michal Polany, 1974). So, modelling of tacit knowledge is a key research area in Artificial Intelligence. Among others, we have selected Ayurvedic domain as a case study for domain with tacit knowledge. They're various reasons for this choice. More importantly, Ayurvedic medicine is full of tacit form of knowledge and we have

This paper describes our research work into the development of a framework for modeling of tacit knowledge. The insight for the construction of the framework is based on the study of classification of individuals in according to Ayurvedic medicine.

2. Ayurvedic Classification of individuals

According to Ayurvedic classification individuals can be grouped into 7 types of their dominance of components such as *Vata*, *Pita*, *Kapha*, *Vata Pita*, *Vata Kapha*, *Pita Kapha*, or *Vata Pita Kapha*. One of the main principles in Ayurvedic medicine is based on the importance of individual differences with regard to treatments. Ayurveda has gone beyond mere classification and identified possible diseases for each category of people. In general population, human constitution is combination of *Vata*, *Pita*, and *Kapha*. Recognition of human constituent in Ayurveda, is currently based on a standard questionnaire on subjective criteria based on ancient theories of Ayurvedic scholar *Charaka*, 1000 BC and *Susruta*, 600 BC. Questions in concerned are very much user-friendly and based on medical theories of Ayurveda, which is used for finding constituent type, has probes such as repeating questions and classification of constituent type. This has been used for classification of individuals for many centuries. There has been no research into improve the questionnaire although people have realised that the classification is not acceptable sometimes. The ayurvedic knowledge of individual classification is a good explain about a model of tacit knowledge.

2.1 Problem definition

In the first place we have tried statistical technique of Principal Component analysis for recognition of any dependencies among classification of individuals. We have used 100 participants for the gathering data. Although the experiment identified several principal components, it was revealed that those components are not significant to consider. However, this decision does not match with the

real world experience, as there were obvious miss match between conclusion through the questionnaire and the actual observations by Ayurvedic physicians. So, it appears that principle component analysis cannot handle the issue we have.

In our research we have used Artificial intelligence techniques for addressing the above issue of recognition of relationships pertaining to tacit knowledge of Ayurvedic classification of individuals. It is well known fact that AI techniques are better at solving real world problems, which cannot be solved otherwise. In particular Fuzzy logic is ideal for handling situations where there are several possibilities in an answer and conclusions are generally vague. Further expert system technology is good at emulation of expert problem solving behaviour and also can be linked up with fuzzy logic. So, we postulate that fuzzy expert systems can produce powerful framework for modelling of domains with tacit knowledge.

3. Framework for modelling tacit knowledge

In general any people who wish to model tacit knowledge can use the framework. However, it can be readily used by Ayurvedic physicians as a consultancy support system. Further, the system can also be used as a teaching tool for Ayurvedic medicine students.

As we mentioned earlier, we have designed and developed a framework for tacit knowledge modelling as a fuzzy-expert system. It is a hybrid intelligent system. Figure 1 shows the top-level design of the system. It consists of Interface, Inference engine, knowledge base, fuzzy logic module, principal component analyser and database. Below is a brief description on these components.

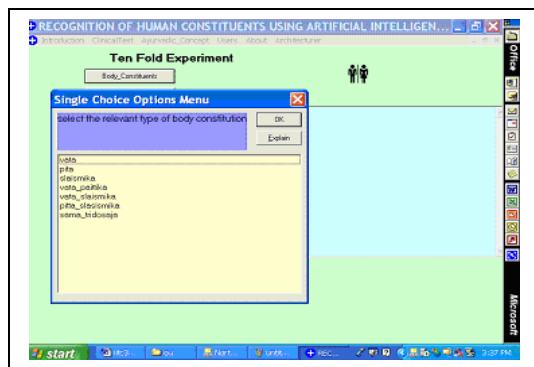


Figure 1: Explanation window of the constituents

3.1 Knowledge base module

The knowledge base contains the domain knowledge useful for problem solving. The knowledge is represented as a set of rules. Each rule specifies a relation and has the IF THEN structure. Rule based is defined using FLEX expert system shell. Further it has been extended to change the rules dynamically. This can be any domain, need not to be the Ayurvedic domain all the time.

3.2 Database module

Tacit knowledge of a particular domain should be stored in the database. The knowledge should be stored in the form of a questionnaire. So, any body wish to model the tacit knowledge can ask some users to answer the questionnaire. The result will also be stored in the database. Principle components of analysed results will also be stored in the database. This has been developed using MS-Access.

3.3 Fuzzy logic module

It has been set a Fuzzy logic module to clarify the final output of the testing module. The Fuzzy logic system has been designed to accept questionnaire results directly and through principal component analysis. Upon entering inputs through this channel, the Fuzzy logic system provides classifications as the output. This has been written using VB.

Fuzzy-expert system

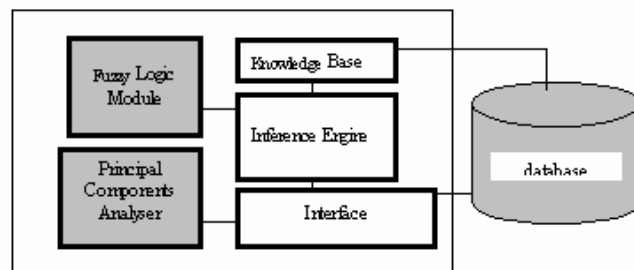


Figure 2 - Top-level design

3.4 Inference engine module

The inference engine module carries out the reasoning whereby the expert system reaches a solution. This can be a slandered inference engine of an expert system. However, additionally, the inference engine in the framework has connection to Principal Component analysis module and Fuzzy logic module.

3.5 Principle component analyser

This module reads from database and get collected data and feed into SPSS. It analyses data with the support from SPSS and send extracted principle component into database. This module implements the principal component analysis as follows.

3.5.1 Number of components

The importance of each PC, in terms of level of data variation explained, is specified by its eigenvalue, the λ term, with $\sum \lambda$ representing the total of the p eigen values. A measure of the proportion of data variation accounted for by each PC, based on the equivalence of eigenvalue and PC variance, is provided by the expression $\lambda / (\sum \lambda)$.

Generally, it is required to select those PCs, which account cumulatively for at least 80% to 90% of the data variation. In addition that each PC must exceed eigenvalue more than 1. However, if nearly all the correlations are less than 0.25, then there is probably not much point in carrying out a PCA. But to reduce even that much of interdependency PCs can be computed.

3.5.2 The interpretation of principal components

Let PC1 be the first principal components, $a_1 \dots a_{72}$ be weightings values, $S_1 \dots S_{72}$ be, Question numbers (variables) related to the symptoms related to Standard Questionnaire.

(variables) related to the symptoms related to Standard Questionnaire.

Let n be no. of principal components extracted.

So PC_i , be no. of principal components. (Where $i = 1 \dots n$)

$$PC_i = a_1 S_1 + a_2 S_2 + \dots + a_{72} S_{72} \quad (2)$$

Further,

$$PC_i = V_i + P_i + K_i \quad (3)$$

Where $i = 1 \dots n$

$$V_i = a_1 S_1 + \dots + a_{24} S_{24} \quad (4)$$

$$K_i = a_{25} S_{25} + \dots + a_{48} S_{48} \quad (5)$$

$$P_i = a_{49} S_{49} + \dots + a_{72} S_{72} \quad (6)$$

For n no. of extracted principal components, following computation is concluded.

$$\text{For Vata: } X = \sum_{i=1}^n V_i \quad (7)$$

According to *Kaiser's* criterion (eigen value >1), principal component extracted, Table: 1.

Question No.	Components			
	PC1	PC2	PC3 PC26
S1	-0.083	-0.072	-0.086	-0.031
S2	0.039	-0.060	-0.061	-0.028
S3	0.067	-0.015	-0.394	-0.031
.
.
.
S72	-0.145	-0.090	0.068	0.110
EigenValue	6.8778	3.7082	3.5941	1.0504
Proportion	0.096	0.052	0.050	0.015

i.e. PC1, PC2,.....PC26

Table 1. Component coefficients

$$\text{For Pita: } Y = \sum_{i=1}^n P_i \quad (8)$$

$$\text{For Kapha: } Z = \sum_{i=1}^n K_i \quad (9)$$

4. Framework in practice

Here we describe how we analyse and model the tacit knowledge domain of Ayurvedic classification of individuals. In the first place we take participants and ask them to answer the questionnaire and store data in the database. The standard questionnaire, which is consisted of 72 multiple-choice questions (MCQ), is used to test human constituent. The questionnaire is user-friendly and easy to understand (Figure 3).

In the pilot study it was found that nearly all the correlations of the questions in concerned are less than 0.25.

In doing so, we have identified 26 principal components in relation to Ayurvedic questionnaire. However, they are not significant to consider according to statistics. This contradicts with real world finding that some questions in the questionnaire have obvious dependencies.

The system now interoperates principal components as Follows.

$$PC1 = -0.083S1 + 0.039S2 + \dots + 0.010S71 - 0.145S72 \quad (10)$$

Further, $PC1 = V1 + P1 + K1$,

$$V1 = -0.0835S1 + 0.039S2 + \dots - 0.080S24 \quad (11)$$

$$K1 = -0.003S25 - 0.011S26 + \dots + 0.267S48 \quad (12)$$

$$P1 = 0.157S49 + 0.084S50 + \dots - 0.145S72. \quad (13)$$

Once the principle components are analysed the system activates the fuzzy logic module. PC will be the input for

Let A be fuzzy set defined in the interval of $[X1 \dots X2]$. . Membership function is as follows.

$$A(X) = \begin{cases} 0 & X \leq X1 \\ (X-X1)/(X2-X1) & X1 < X < X2 \\ 1 & X \geq X2 \end{cases}$$

$$V \quad A(X) = \begin{cases} 0 & X \geq -5.8405 \\ (X + 5.8405) / (-29.2025) & -43 < X < -5.8405 \\ 1 & X \leq -35.043 \end{cases} \quad (17)$$

Fuzzy logic module has been set up with principal component (PC). Following 3 Fuzzy sets are used to represent the constituents of *Vata*, *Pita* and *Kapha*. Let V, P, and K are the linguistics variables for defined Fuzzy sets. Determining the bound values can be computed as follows.

Lower bound value is obtained when Marks = 1 (each question) for each constituent type and upper bound value is obtained when Marks = 6 (each question) for each constituent respectively.

Let PC1 be the first principal component, then according to proven concept in design module Likewise other 26 principal components can also be shown.

So Final computational procedure is as follows,

$$\text{For } Vata: X = \sum_{i=1}^{26} V_i \quad (14)$$

$$\text{For } Pita: Y = \sum_{i=1}^{26} P_i \quad (15)$$

$$\text{For } Kapha: Z = \sum_{i=1}^{26} K_i \quad (16)$$

the fuzzy logic module. Now the fuzzy logic module generates the membership function as follows.

$$P(Y) = \begin{cases} 0 & Y \geq -294.1801 \\ (Y + 294.1801) / (-1480.9009) & -1777.081 < Y < -294.1801 \\ 1 & Y \leq -1777.081 \end{cases} \quad (18)$$

$$K(Z) = \begin{cases} 0 & Z \leq 112.5047 \\ (Z - 112.5047) / 562.5235 & 112.5047 < Z < 675.0282 \\ 1 & Z \geq 675.0282 \end{cases} \quad (19)$$

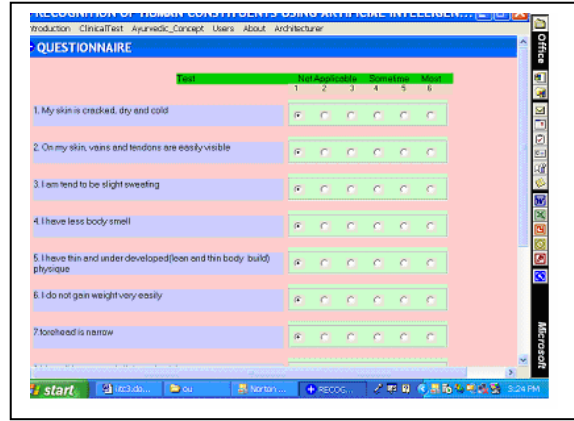


Figure 3: Analysis window of the Questionnaire

Fuzzy rules are as follows,

Rule 1: If $X \geq -5.8405$
then $VATA = 0 \%$

Rule 2: If $-35.043 < X < -5.8405$
then $VATA = ((X + 5.8405) / (-29.2025)) * 100 \%$

Rule 3: If $X \leq -35.043$
then $VATA = 100 \%$

Rule 4: If $Y \geq -294.1801$
then $PITA = 0 \%$

Rule 5: If $-1777.081 < Y < -35.043$
then $PITA = ((Y - 294.18) / (-29.220)) * 100 \%$

Rule 6: If $Y \leq -1777.081$
then $PITA = 100 \%$

Rule 7: If $Z \leq 112.5047$
then $KAPHA = 0 \%$

Rule 8: If $112.5047 < Z < 675.0282$
then $KAPHA = ((Z - 112.5047) / 562.5235) * 100 \%$

Rule 9: If $Z \geq 675.0282$
then $KAPHA = 100 \%$

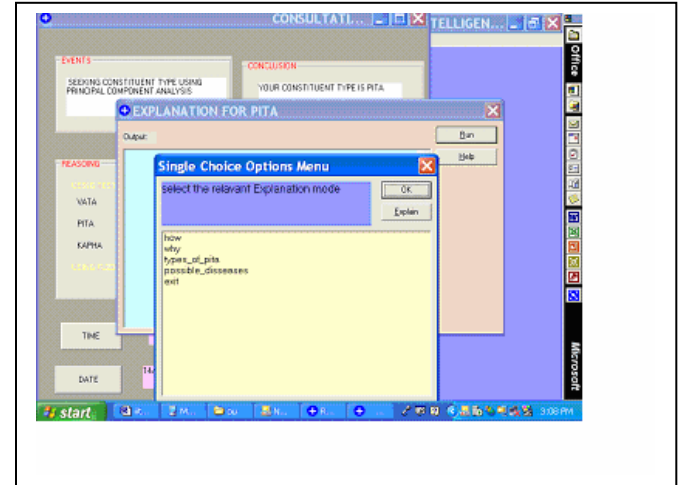


Figure 4: Explanation window of the Expert system

5. Conclusion

Scope of this research was to study the power of intelligent techniques that have been given promising results when classical methods fail for enhancing the method of human constituent classification in Ayurvedic Medicine. Knowledge is the essential tool for a person or an organization. Real world applications normally belong to domains with tacit knowledge. Explicit knowledge is required for increasing the accuracy of the

output. To model the inference engine mechanism, it has been used the principal component analysis and Fuzzy Logic. It has been develop generalized framework

For modeling` f tacit knowledge, it has been integrated the tool within standard inference engine of expert system shell.

Ayurvedic students can use this system as a learning module and also recognition of constituents. Though the system is developed very satisfactory, there are further enhancements to be done to the system, to implement as a web application where clients can log on to the system rather than using as stand-alone system.

References

- [1] Karunananda A.S, “How To Do Research”, 2000.
- [2] Jonson L, “Expert system Architectures”, Kopan Page Limited, 1988.
- [3] George J, “Fuzzy sets and Fuzzy Logic”, Prentice Hall of India, 1997.
- [4] Noak V, “Discovering the world with Fuzzy logic”, A Springer – Verlag Company, 2000, PP. 3 – 50.
- [5] Dubey G.P, “The Physiological concepts in Indian medicine”, Science and Philosophy of Indian medicine, 1978.
- [6] Tripathi S.N, “Clinical Diagnosis”, Science and Philosophy of Indian medicine, 1978.
- [7] Chatfield C,” Introduction to Multivariate Analysis”, Chapman and Hall, 1996.
- [8] Morrison D.F, “Multivariate Statistical Methods”, Mc Graw-Hill, Inc. 1990, PP. 312 - 357.
- [9] Ross Dawson, , “Developing Knowledge-Based Client Relationship ”, Butterworth Heinemann, 2001
- [10] Richards D. and Bush P., “Measuring, Formalizing and Modeling Tacit Knowledge” IEEE/Web Intelligence Conference (WI-2003) Beijing.
- [11] M. Dzbtor, Explication of Design Requirements Through Reflection on Solutions. In Proc. of 4th IEEE Conference. on Knowledge-based Intelligent *Engineering Systems*. Brighton, UK, pages 141-144, 2000.



Image Coding Using the Self-Similarity of Wavelet High-Frequency Components

S.Selvarajan

N.D.Kodikara

Dept. of Physical Science
Vavuniya Campus of the University of Jaffna
Station Road
Vavuniya 43000
E-Mail:ksselvarajan@vavu.jfn.ac.lk

Dept. of Comm. & Media Tech.
University of Colombo
School of Computing
Colombo-03
E-Mail:ndk@ucsc.cmb.ac.lk

Abstract

In this paper a novel approach Self-Similar Wavelet High-Frequency Components coding (SWHFC) to compress an image is proposed, based on the self-similarity of wavelet high-frequency components. The self-similarity of wavelet transformed image is explored by using the multiresolution concept of weighted finite automata (WFA). This scheme combines the multiresolution concept of WFA and wavelet multiresolution concept by introducing a predicative coefficients along with the each weights of WFA. Objective of this is to predict wavelet coefficients across scales, as to reduce the number of wavelet coefficients to be coded and the corresponding information of those.

Keywords: Image compression, wavelets, MRA, WFA, image smoothness, linear and nonlinear approximation

1. Introduction

In lossy image compression, two types of schemes are in use: fractal and transform coding. Fractal coding based on IFS, first proposed by Jacquin [7]. WFA is the generalization of IFS and proposed by Culik [1]. Transform coding scheme uses wavelet transform as an alternative to the DCT used earlier in JPEG.

In image processing most of the images are typically represents spatial trends, or areas of high statistical spatial correlation. However, anomalies, such as edges or object boundaries, take on perceptual significance that is far greater than their numerical energy contribution to an image. The basic idea of wavelet transform is to represent any arbitrary function f as a superposition of wavelets. Any such superposition decomposes f into different scale levels, where each level is then further decomposed with a resolution adapted to the level. One way to achieve

such a decomposition writes f as an integral over a and b of $\psi_{a,b}$ with appropriate weighting coefficients. Store only the relevant frequency components known as wavelet coefficients. A major difficulty is that fine detail coefficients representing possible anomalies constitute the large number of coefficients and therefore to make effective use of the multiresolution representation, much of the information is contained in representing the position of those few coefficients corresponding to significant anomalies.

The technique of this research allow coders to effectively use the power of multiresolution representation by efficiently representing the positions of the wavelet coefficients representing significant anomalies. A notable breakthrough was the introduction of Embedded Zero-tree Wavelet (EZW) coding by Shapiro [6]. In view of this context, it has been proposed to code natural images by nonlinear estimation of wavelet coefficients and combining both WFA and EZW by bounding the error incurred by quantizing wavelet transform coefficients and coding in WFA.

In this paper we present a recent result in combining those two, using the multiresolution properties of these two schemes. Culik [2] proposed a method to combine theses two scheme for fractal and smooth images, it reconstruct the image from the directly from WFA, where coefficients are coded. Our scheme differs from that, as it predict the coefficients by using the multiresolution concept of both and hierarchical nature of wavelet the transform. It is also applicable to natural images as well.

2. Weighted Finite Automata (WFA)

WFA is introduced by Culik as a device to compress images from the pixels grey-scale value. Later it was studied by Hafner [3]. The earlier is known as linear automata and

the later is known to be hierarchical because of its nature. We use the second one in our algorithm, this is somewhat similar to Laplacian Pyramidal coding [8].

Formal definition of the WFA is given [1], [3]. Wavelet transformed image represented in Mallat [4] form could be understandable as an independent image at different resolution in their corresponding orientation, i.e., horizontal, vertical and diagonal. These images (high-frequency components) are similar to each other by a the scaling factor with respect to their orientation. So that, coding the each highest-frequency component and low-frequency components are suffice to reconstruct the image by using the prediction across the scales of wavelet transform.

3. Image Compression

By an image, it means a digitized grey-scale picture, that $2^m \times 2^m$ pixels, $m \in \mathbb{Z}$ each of which takes a value p_j such that, $0 \leq p_j \leq 2^n - 1$, where $j = (j_1, j_2)$, j_1 and j_2 are index of rows and columns respectively. In this analysis an image is described as a function f on a unit square $I = [0, 1]^2$,

$$f(x) = p_j \quad \text{for} \quad \frac{j_1}{2^m} \leq x_1 \leq \frac{j_1 + 1}{2^m} \quad \text{and} \quad \frac{j_2}{2^m} \leq x_2 \leq \frac{j_2 + 1}{2^m}$$

where $x = (x_1, x_2)$ in I

By applying discrete transformation, e.g., DCT, Haar transform, we have

$$f = \sum_{k \geq 0, j = (j_1, j_2)} c_{j,k} \phi_{j,k} \quad (1)$$

where $c_{j,k}$ are coefficients as the result of applying discrete transformation on the basis function $\phi_{j,k}$.

Now the image compression problem viewed as the result of approximating f by a second (compressed) function \tilde{f} , i.e., given the transform, the algorithm then calculates the quantized coefficients $\tilde{c}_{j,k}$ and the compressed function takes the form as:

$$\tilde{f} = \sum_{k \geq 0, j = (j_1, j_2)} \tilde{c}_{j,k} \phi_{j,k} \quad (2)$$

4. Wavelet Representation of an Image

This section describes theoretical property of how an image is viewed as an image itself in wavelet transformation. Also, the principle underlying behind the coding of wavelet coefficients in par with pixels being coded with WFA.

A digitized image is that the pixel value (observed) are samples, which depend on measuring device of the intensity field $F(x)$ for x on the square $I = [0, 1]^2$. Pixel samples are well modeled by averaging the intensity function F over all squares.

Assume that 2^{2m} pixel values p_j are indexed by $j := (j_1, j_2)$, $0 \leq j_1, j_2 \leq 2^{2m}$ of 2^{2m} rows and columns and that each measurement is the average value of F on the subsquare covered by that pixel. To fix this notation, the j^{th} pixel covers the square $I_{j,m}$ with side length 2^{-m} and lower left corner at the point $j/2^m$. Denote the characteristic function of I by $\chi = \chi_I$ and $L_2(I)$ -normalized characteristic function of $I_{j,m}$ by $\chi_{j,m} = 2^m \chi_{I_{j,m}} = 2^m \chi(2^m \cdot - j)$. Then the pixel value would be as:

$$p_j = 2^{2m} \int \chi(2^m x - j) F(x) dx \\ = 2^m \langle \chi_{j,m}, F \rangle$$

The standard practice in wavelet-based image processing is to use the observed pixel value p_j to create the function.

$$f_m = \sum_j p_j \chi(2^m \cdot - j) \\ = \sum_j \langle \chi_{j,m}, F \rangle \chi_{j,m}$$

which is known as observed image. If the wavelet expansion of the intensity field F is

$$F = \sum_{0 \leq k} \sum_{j, \psi} c_{j,k,\psi} \Psi_{j,k}$$

then we have

$$f_m = \sum_{0 \leq k \leq m} \sum_{j, \psi} c_{j,k,\psi} \Psi_{j,k}$$

Note that f_m is the $L_2(I)$ projection of F onto $\text{span}\{\chi_{j,m}\}_j = \text{span}\{\Psi_{j,k}\}_{0 \leq k \leq m, j, \psi}$.

4.1. Multiresolution Concept

In computer vision, it is difficult to analyze the information content of an image directly from the grey-level intensity of the image pixels. The size of the neighborhood where the contrast is computed must be adapted to the size of the objects. The size define the resolution of reference for measuring the local variation of the image. Given a sequence of increasing resolutions $(r_j)_{j \in \mathbb{Z}}$, the details of an image at the resolution r_j are defined as the difference of information between its approximation at the resolution r_j

and its approximation at the lower resolution r_{j-1} .

Representing a function by a array of pixels smooths out details smaller than a pixel. But, because the scale of observation is arbitrary, one pixel can represent an area of any size. To avoid this problem, scale independent representation should be developed. This could be achieved through the use of multi resolution analysis. Intuitively, a MRA is a collection of subsets V_i of L_2 , $i \in \mathbb{Z}$. Each V_i contains all functions whose details smaller than $(r_i)_{i \in \mathbb{Z}}$ is removed. Removal of details depends upon the particular type of MRA.

The Laplacian pyramid data structures suffer from the difficulty that data at separate levels are correlated. There is no clear model which handles this correlation. It is thus difficult to know whether a similarity between the image details at different resolutions is due to the property of the image itself or to the intrinsic redundancy of the representation. Further, this does not introduce any spatial orientation selectivity into the decomposition process.

Pyramidal implementation have been developed for computing the multiresolution transform based on convolution with quadrature mirror filters. The signal can be reconstructed by reversing the above process. A multiresolution transform also decomposes the signal into a set of frequency channels of constant bandwidth on logarithmic scale. It can be interpreted as a discrete wavelet transform.

4.2. Multiresolution Representation of an Image

For computational reason and self-similarity of the space-frequency plane of sub-bands, the Haar transform of the representation of f is considered. Let

$$\Psi(x) = \begin{cases} -1 & 0 \leq x < 1/2 \\ 1 & 1/2 \leq x < 1 \end{cases} \quad (3)$$

and

$$\phi(x) = 1 \quad 0 \leq x < 1 \quad (4)$$

The four basis function for the local Haar representation of functions for 2-D are:

$$\psi^1(x, y) = \phi(x)\psi(y) \quad \psi^2(x, y) = \psi(x)\phi(y)$$

$$\psi^3(x, y) = \psi(x)\psi(y) \quad \psi^4(x, y) = \phi(x)\phi(y)$$

By dilation and translation, we have $\psi_{j,k}^{(i)} = \psi^{(i)}(2^k \cdot - j)$ $i = 1, \dots, 4$, $j \in \mathbb{Z}^2$ $k \geq 0$ Then

$$f = P_{(2j_1, 2j_2), k} \Phi_{(2j_1, 2j_2), k} + \dots + P_{(2j_1+1, 2j_2+1), k} \Phi_{(2j_1+1, 2j_2+1), k} \quad (5)$$

where $p_{j,k}$ is the grey-scale of a pixel

$$f = \frac{c_{j,k-1}^{(1)}}{4} \Psi_{j,k-1}^{(1)} + \dots + \frac{c_{j,k-1}^{(4)}}{4} \Psi_{j,k-1}^{(4)} \quad (6)$$

The coefficient $c_{j,k-1}^{(i)}$ are determined by the following identity matrix:

$$\begin{pmatrix} c_{j,k-1}^{(1)} \\ c_{j,k-1}^{(2)} \\ c_{j,k-1}^{(3)} \\ c_{j,k-1}^{(4)} \end{pmatrix} = \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} P_{(2j_1, 2j_2), k} \\ P_{(2j_1+1, 2j_2), k} \\ P_{(2j_1, 2j_2+1), k} \\ P_{(2j_1+1, 2j_2+1), k} \end{pmatrix} \quad (7)$$

all the $p_{j,k}$ are integers, then so are the $c_{j,k-1}^{(i)}$. With this transform we have:

$$f = \frac{1}{4} \sum_{k \geq 0} \sum_{j \in \mathbb{Z}^2} \sum_{i=1}^4 c_{j,k-1}^{(i)} \Psi_{j,k-1}^{(i)} \quad (8)$$

4.3. Similarities Across Scales

Wavelet coefficients associated with high-frequency components depends on a small rough region located around the edge and not on of overall smoothness of the picture. Conversely, smooth regions are less likely to be rippled due to nearby edge. The projection of a function f on two subspaces V_i and V_{i+1} , similarities between them are observed. These two projections isolates two different frequency sub-bands. From the previous fact, if function $A_i f$ is rough in some regin, $A_{i+1} f$ is also rough in the corresponding region. Our aim is to explore the nature of this similarity between the frequency sub-bands.

4.4. Linear and Nonlinear Approximation

Let $f(x) = p_j$ for $x \in I_{j,m}$ (support of the image I defined as earlier)

$$F = \sum_{k \geq 0} \sum_{j,k} c_{j,k} \Psi_{j,k} \quad (9)$$

$$f_m = \sum_{0 < k < m} \sum_{j,k} c_{j,k} \Psi_{j,k} \quad (10)$$

where each pixel p_j , $j \in \mathbb{Z}_m^2$, is the average of the intensity $F(x)$ on $I_{j,m}$ which is defined to be the square of side length 2^{-m} and lower left corner located at $j/2^m$. Consider the smoothness of any order $0 < \alpha < \beta$. The $L_2(I)$, $W^\alpha(L_2(I))$, $0 < \alpha < 1/2$ and $B_q^\alpha(L_q(I))$, $0 < \alpha < 1$ and $q = 2/(1 + \alpha)$, norms of f are bounded by the corresponding norms of F . Therefore, linear approximation \tilde{f} of f is given by:

$$f_K = \sum_{0 < k < K} \sum_{j,k} c_{j,k,\psi} \psi_{j,k} \quad (11)$$

where K is the smallest integer such that $\lambda 2^{2\beta k} \geq 1$. This means, approximating all the coefficients $c_{j,k,\psi}$ with frequency less than 2^K , $K \leq m$.

In nonlinear approximation, fix a number $n \geq 0$ and approximate f by

$$f_\lambda = \sum_{\lambda \in \Lambda_0} c_\lambda \phi_\lambda \quad (12)$$

where Λ_0 is an arbitrary subset of Λ with 2^n elements. The best nonlinear approximation is obtained by taking Λ_0 as the set of the 2^n indices λ for which $|c_\lambda|$ is the largest. With this process the image is smoothen, just as convolving with a smoothing kernel.

5. Wavelets, Quadtrees and WFA

Because of the dyadic nature of the wavelet decomposition, the 2-D wavelet transform is typically arranged in three subbands, corresponding to their orientation of the wavelet basis functions. Each subband can be organized into a quadtree, as described below.

In the quadtree interpretation of the 2-D wavelet transform, each node i is labeled with a wavelet coefficient $c_{i,j,\psi^{(k)}}$, where the corresponding wavelet basis function $\psi^{(k)}$ has approximate support on a square, dyadic block B_i , in the image. The width of this block is given by $M = 2^{-l}N$, where l is the depth of node i in the quadtree and N is the width in pixels of the square image (assumed to be the power of two).

Except at the finest level, each node has four children representing $M/2 \times M/2$ dyadic blocks that combine to tile the same $M \times M$ image block as their parent. Objective is to code the finest level coefficients from which estimate the next lower high-frequency subband components, in order to reduce the coding, position information of each wavelet coefficients.

Images are transformed using Haar basis orthogonal function. Each sub-band is assumed as the independent images and performed into horizontal, vertical and diagonal groups. The correlation between these groups of image components are very high, i.e., wavelet transform by Haar basis orthogonal function carries out the mean value among adjacent pixels can decompose into the space-frequency. Exactly, this process repeats the finite difference operation after averaging the adjacent pixels.

This could be seen as, higher frequency component of each frequency sub-band, wavelet coefficients in quad-tree form (2×2) having as the initial distribution and each

edge having the weight $w_{m,n}$. It could be repeated for finite number of times to infer the wavelet coefficients in the corresponding lower frequency sub-bands. See figure 1.

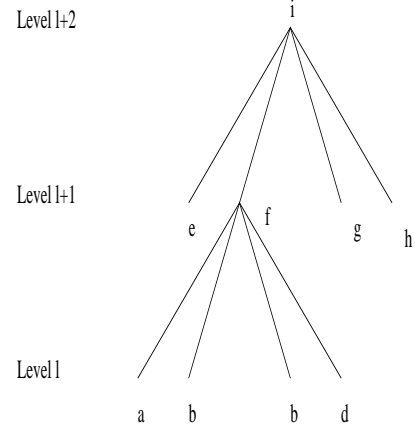


Figure 1: Coefficients construction.

5.1. Defining an Automata for Higher-frequency Components

It is noted high-frequency components of an image are similar across scales. It is obvious to define a WFA to represent such similarities. Since the WFA is the generalization of IFS. Also, wavelet transform consists hierarchical nature and multiresolution property. These could be incorporated in WFA. These properties ensures the multiresolution property of an image high frequency components, i.e., a function at resolution 2^k can be computed from the resolution 2^{k+1} by averaging the adjacent 2×2 coefficients. On the basis of above concept define this representation as follows:

Definition 1 Let $c_{l,\psi}(2i+m, 2j+n) \neq 0$ be the coefficient of highest-frequency components and $w(m,n)$ be the weight of each edge that correspond to next lower high-frequency component coefficient in the same orientation, for levels $0 < l < N$ and $0 \leq i, j < 2^k$, $k \in \mathbb{Z}$, then the coefficient of the next high-frequency component in the same orientation is estimated as:

$$\hat{c}_{l+1,\psi}(i^{l+1}, j^{l+1}) = \sum_{m=0}^3 \sum_{n=0}^3 w(m,n) \tilde{c}_{l,\psi}(2i+m, 2j+n) \quad (13)$$

where $w(m,n) = 1/2$ or 1 (scale invariant coefficient).

6. Zerotree for Smooth Regions

The smoothness of wavelet is often characterized by the number of vanishing moments. A function f defined over an interval $[a, b]$ is said to have n vanishing moments if and only if

$$\int_a^b f(x)x^i dx = 0 \quad \text{for } i = 0, 1, \dots, n-1 \quad (14)$$

It refers to the decay of wavelet coefficients through the scales. Intuitively, we expect smooth image regions will have small wavelet coefficients.

Suppose node i has support on a dyadic image block B_i , that is characterized as smooth. Because all nodes descending from i can also be characterized as smooth, this model assumes $c_{i,j,\psi^{(k)}} = 0$, as such corresponding coefficients in the next frequency subband component is zero. It is simply a fixed approximation to the wavelet coefficients. This tree-structured approximation is known as a zero-tree [6]. A combination of zerotree, nonlinear approximation and prediction of wavelet coefficients through the scale of high-frequency sub-band components are suffice for image coders.

A notable breakthrough in wavelet based coding was the introduction of Embedded Zerotree Wavelet (EZW) coding by Shapiro [6]. A significance map was defined as an indication of whether a particular coefficient was zero or nonzero (i.e., significant) relative to a quantization level. Defining a wavelet coefficient as insignificant with respect to a threshold T if $|c_{i,j}| < T$, the EZW algorithm hypothesized that if a wavelet coefficient at coarse scale is insignificant with respect to a given threshold T , then all wavelet coefficients of the same orientation in the same spatial location at finer scales are likely to be insignificant with respect to T .

- if the corresponding coefficient in the next lower frequency sub-band is insignificant, then all the four coefficients are insignificant regardless of their values
- similarly, if the coefficients produce insignificant coefficient value to next lower frequency sub-band with our algorithm, then all four coefficients are insignificant

7. Coding and Decoding Algorithm

Images are transformed in three down sampling, using Haar basis orthogonal function. Haar basis function carries out the mean value among the adjacent pixels can decompose into the space-frequency. Exactly, this process repeats the finite difference operation after averaging the adjacent pixels. So that, it could be viewed as, higher-frequency components are in quad- tree form 2×2 as the initial distribution and each having the weight $w_{n,m}$ may be $1/2$ or 1 . Usually the weight is $1/2$, in case of scale invariant coefficients the weight is 1 . With this terminology, the corresponding next lower high-frequency sub-band coefficients could be inferred from (13). This process could be

repeated for finite number of times to infer the coefficients of high-frequency sub-band components.

Encoding Algorithm:

1. Given image is transformed using Haar orthogonal basis filter in three dyadic levels and wavelet coefficients are represented in Mallat form
2. Coefficients are quantized by using nonlinear estimation
3. Lower frequency sub-band is transmitted/stored
4. Highest frequency sub-band of each components are quantized with zero-tree wavelet coefficient concept introduced in this scheme
5. Each, nonzero coefficients of highest frequency sub-bands are (2×2) are coded with quad-tree address and transmitted/stored
 - 5.1 scale invariant coefficients are coded with their address
 - 5.2 other non-zero coefficients are coded with their address

Next how the weights are determined. Usually weight of the coefficients is half, but some coefficients are scale invariant in this case the weight is one. These weights are determined as follows:

```
if (c_{i,0} + ... + c_{i,3}) < threshold
w_{i,j}=1
else
w_{i,j}=1/2
```

Decoding Algorithm:

1. Lower frequency component coefficients are placed in corresponding array position
2. Highest frequency sub-bands coefficients of scale invariant are placed in the corresponding position of the array
3. Next high frequency sub-bands are estimated recursively: each coefficients having weight one; corresponding coefficient is the summation of initial distribution (coefficients)
4. Next high frequency sub-bands (other coefficients that are not estimated from the previous step) are estimated recursively: each coefficients having weight $1/2$; corresponding coefficient is the half of the summation of initial distribution (coefficients)
5. Inverse wavelet transform is obtained to reconstruct the image with the array

Method	CR	RMSE	PSNR(dB)
SWHFC	0.03bpp	14.05	25.17
SQS [12]	0.036bpp	-	25.86
WFC [13]	0.036bpp	-	26.42
SPIHT [14]	0.036bpp	-	26.49
FZW [11]	0.036bpp	-	26.49
Spatial fractals [7]	17.4	-	24.90
DCT fractals [10]	18.5	-	26.10
JPEG	0.15bpp	-	26.44

Table 1: Comparison of algorithms for Lena image 512×512

8. Conclusion

Inorder to evaluate the results root mean squared (RMS) error and peak-to-peak signal-to-noise ratio (PSNR) values are used against to the number of nonzero wavelet coefficients. An image Lena 512×512 is used as the test image. The high-frequency nonzero wavelet coefficients are coded with SWHFC and the quantized coefficients from the same scale are used to reconstruct the image as with linear approximation. The results are compared with other methods in the table 1.

The other methods such as SPIHT, FZW etc., give slightly better results than SWHFC, but advantages of this method remains as with the WFA image coding over the fractal image coding. Also, it describe an novel algorithm to encode an image wavelet coefficients with the WFA coding efficiently. This is a significant improvement in coding wavelet coefficients with WFA.

It is obvious that SWHFC gives high compression at a given fidelity. The coded coefficients with this method may be further coded with an entropy coder to obtain further compression. Fidelity may increase with increasing the non zero wavelet coefficients and/or progressive fidelity transmission.

Haar orthonormal basis function is chosen as the wavelet transform, since the images with edges give high-frequency coefficients than the smooth wavelets, as this method depends on the self-similarity between the high-frequency components. With the increasing threshold value or reducing the number of coefficients, increase in tiling effect. Also, WFA coding of pixels introduces blocking artifact, but this effect is greatly reduced as the wavelet coefficients are coded. This is the obvious advantage in using wavelet coefficients with WFA rather than pixel coding with WFA. It is a better approach than the linear approximation.

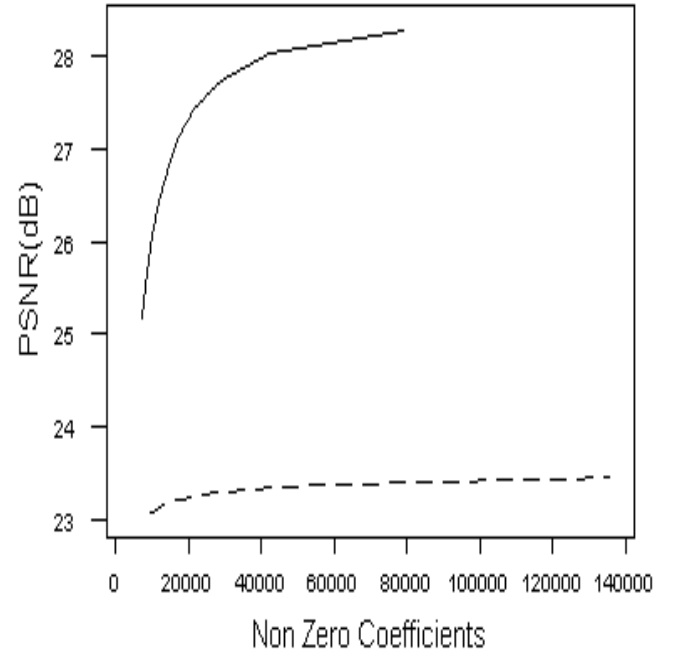


Figure 2: PSNR versus No. of non-zero coefficients for our algorithm (—) and linear approximation (---).

References

- [1] K.Culik II and J.Kari, *Image Compression Using Weighted Finite Automata*, Comput. & Graphics, 17, 3 pp. 305-313 (1993)
- [2] K.Culik II and S.Dube, *Implementing Daubechies wavelet transform with weighted finite automata*, Acta Informatica 34, 347-366(1977)
- [3] U.Hafner, *Asymmetric Coding in (m)-WFA Image Compression*, preprint
- [4] S.G.Mallat, *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*, IEEE Trans. on Pattern Analysis and Machi. Intell., vol. 11, 7, July 1989
- [5] R.A.DeVore, B.Jawerth and B.J.Lucier, *Image Compression through Wavelet Transform coding*, IEEE Trans. on Information Theory 38, 719-746(1992)
- [6] J.Shapiro, *Embedded image coding using zerotrees of wavelet coefficients*, IEEE Trans. Signal Processing, vol. 41, pp. 3445-3462, Dec. 1993
- [7] A.E.Jacquin, *A Novel Fractal Block-Coding Technique for Digital Images*, Proc. ICASSP, pp. 2225-2228, 1990

- [8] P.J.Burt and E.H.Adelson, *The Laplacian Pyramid as a Compact Image Code*, IEEE Trans. on Comm., vol 31, 4, Apr. 1983, pp. 532-535
- [9] Y.Ueno, *Wavelets and fractal image compression based on their self-similarity of the space-frequency plane of images*, LNCS 2251, pp. 87-98, 2001, Springer-Verlag, Berlin.
- [10] Y.Zhao and B.Yuon, *Image Compression Using Fractals and Discrete Cosine Transform*, Electronic Letters, 17th March 1994, vol. 30, No 6, pp. 474-475
- [11] T.Kim, R.E.Van Dyck and D.J.Miller, *Hybrid Fractal Zerotree Wavelet Image Coding*, Signal Processing Image Communication, 2002 <http://w3.antd.nist.gov/pubs02.shtml>
- [12] G.Davis, *A Wavelet-Based Analysis of Fractal Image Compression*, IEEE Trans. Image Processing, vol 7, 2, Feb. 1998, pp. 141-154
- [13] J.Li and C.C.Kuo, *Image Compression with Hybrid Wavelet-Fractal Coder*, IEEE Trans. Image Processing, vol 8, 6, June 1999, pp. 868-874
- [14] A.Said and W.A.Pearlman, *A New, Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees*, IEEE Trans. on Circuits and Systems for Video Tech., vol 6, 3, June 1996 243-250



Figure 3: Lena original image 512×512 and Quantized Haar coefficients.

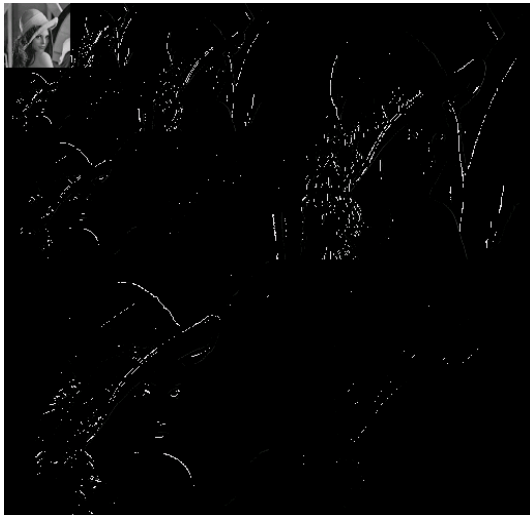


Figure 4: Estimated coefficients and Reconstructed image 25.17dB at 0.03bpp 512×512 .



Comparative Study on Security Issues of Wireless LAN

Md. Enamul Haque, Meeraz Ahmed Saadi,
 Dept. of Computer Science and Mathematics, Bangladesh Agricultural University,
 Mymensingh-2202, Bangladesh
 Dept. of Computer Science and Engineering, East West University, Dhaka-1212, Bangladesh.
 Email: nico@bttb.net.bd, amisaadi@dhaka.net

Abstract

Wireless local area network (WLANs) is popular in both individual and corporate level. But it is less lack of any viable security mechanism. Although IEEE proposed long term security architecture for WLAN which is called Robust Security Network (RSN) still there are flaws in several basic design. In this paper we present some security risks of wireless LAN and its possible solutions.

Keywords : Wireless LAN (WLAN), Robust Security Network (RSN), Intrusion Detection system (IDS), Access Point, Radio Frequency, Service Set Identifier, Wired Equivalent Privacy, MAC address.

1. Introduction

Wireless LAN technology is an IEEE 802.11 standard which is a flexible data communication system implemented as an extension to or as an alternative for, a wired LAN within an area. Using electromagnetic waves or Infrared or Direct sequence spread spectrum, WLANs transmit and receive data over the air, minimizing the need for wired connect. As there is no wire it is hard to monitor its security at every second. First we focused on how the usual Network security works.

2.1 Network Security

Network security is a security measures that are needed to protect data during their transmission, and to guarantee that data transmissions are authentic. However, Network operating systems provide basic security features, such as user identification and authentication, probably by password. Sophisticated Network systems can permit network supervisors to assign varying access rights to individual users. All users, for example, could access word processing software, but only certain users could access payroll files. Some network software can limit

how many times users can call up a particular file and generate an audit trail of who looked at which files.

2.2 Need for the Network Security

In the ever-changing world of global data communications, inexpensive Internet connections, and fast-paced software development, security is becoming more and more of an issue. Security is now a basic requirement because global computing is inherently insecure. As user's data goes from point A to point B which may pass through several other points along the way, giving other users the opportunity to intercept, and even alter, it. Even other users on the system may maliciously transform the data into something the user did not intend. Unauthorized access to the system may be obtained by intruders, also known as "crackers", who then use advanced knowledge to impersonate one, steal information from one, or even deny anyone access to one's own resources.

2.3 Security Requirements

Computer and network security address three types of security requirements:

1. Secrecy. Requires that the information in a computer system only be accessible for reading by authorized parties.
2. Integrity. Requires that only authorized parties can modify computer system assets.
3. Availability. Requires that computer system assets are available to authorized parties.

2.4 Security Systems Wireless LAN Solutions

There are numerous methods available to exploit the security of wired networks via wireless LANs. Layered security and well thought out strategy are necessary steps to locking down the network. Applying best practices for wireless LAN security does not alert the security manager or network administrator when the security has been

compromised. Intrusion Detection Systems (IDS) are deployed on wired networks even with the security provided with VPN and firewalls. However, wire-based IDS can only analyze network traffic once it is on the wire. Unfortunately, wireless LANs are attacked before entering the wired network and by the time attackers exploit the security deployed, they are entering the network as valid users. For IDS to be effective against wireless LAN attacks, it first must be able to monitor the airwaves to recognize and prevent attacks before the hacker authenticates to the AP.

2.5 Principles of Intrusion Detection

Intrusion Detection is the art of detecting inappropriate, incorrect, or anomalous activity and responding to external attacks as well as internal misuse of computer systems. Generally speaking, IDS are comprised of three functional areas:

- A stream source that provides chronological event information.

- An analysis mechanism to determine potential or actual intrusions

- A response mechanism that takes action on the output of the analysis mechanism.

In the wireless LAN space, the stream source would be a remote sensor that promiscuously monitors the airwaves and generates a stream of 802.11 frame data to the analysis mechanism. Since attacks in wireless occur before data is on the wired network, it is important for the source of the event stream to have access to the airwaves before the AP receives the data. The analysis mechanism can consist of one or more components based on any of several intrusion detection models. False positives, where the IDS generated an alarm when the threat did not actually exist, severely hamper the credibility of the IDS. In the same light, false negatives, where the IDS did not generate an alarm and a threat did exist, degrade the reliability of the IDS.

Signature-based techniques produce accurate results but can be limited to historical attack patterns. Relying solely on manual signature-based techniques would only be as good as the latest known attack signature until the next signature update. Anomaly techniques can detect unknown attacks by analyzing normal traffic patterns of the network but are less accurate than the signature-based techniques. A multi-dimensional intrusion detection approach integrates intrusion detection models that combine anomaly and signature-based techniques with policy deviation and state analysis.

2.6 Vulnerability Assessment

Vulnerability assessment is the process of identifying known vulnerabilities in the network. Wireless scanning tools give a snapshot of activity and identify devices on each of the 802.11b channels and perform trend analysis to identify vulnerabilities. A wireless IDS should be able to provide scanning functionality for persistent monitoring of activity to identify weaknesses in the network.

The first step in identifying weakness in a Wireless LAN deployment is to discover all Access Points in the network. Obtaining or determining each one's MAC address, Extended Service Set name, manufacturer, supported transmission rates, authentication modes, and whether or not it is configured to run WEP and wireless administrative management. In addition, identify every workstation equipped with a wireless network interface card, recording the MAC address of each device.

The information collected will be the baseline for the IDS to protect. The IDS should be able to determine rogue AP's and identify wireless stations by vendor fingerprints that will alert to devices that have been overlooked in the deployment process or not meant to be deployed at all.

Radio Frequency (RF) bleed can give hackers unnecessary opportunities to associate to an AP. RF bleed should be minimized where possible through the use of directional antennas discussed above or by placing Access Points closer to the middle of buildings as opposed to the outside perimeter.

2.7 Defining Wireless LAN Security Policies

Security policies must be defined to set thresholds for acceptable network operations and performance. For example, a security policy could be defined to ensure that Access Points do not broadcast its Service Set Identifier (SSID). If an Access Point is deployed or reconfigured and broadcasts the SSID, the IDS should generate an alarm. Defining security policies gives the security or network administrator a map of the network security model for effectively managing network security.

With the introduction of Access Points into the network, security policies need to be set for Access Point and Wireless Station configuration thresholds. Policies should be defined for authorized Access Points and their respective configuration parameters such as Vendor ID, authentication modes, and allowed WEP modes. WEP algorithm is a part of 802.11 standard which may be in 40 bit or 128 bit version. Allowable channels of operation and normal activity hours of operation should be defined for each AP. Performance thresholds should be defined for minimum signal strength from a wireless station

associating with an AP to identify potential attacks from outside the building.

The defined security policies form the baseline for how the wireless network should operate. The thresholds and configuration parameters should be adjusted over time to tighten or loosen the security baseline to meet real-world requirements. For example, normal activity hours for a particular AP could be scaled back due to working hour changes. The security policy should also be changed to reflect the new hours of operation. No one security policy fits all environments or situations. There are always trade offs between security, usability and implementing new technologies.

2.8 State-Analysis

Maintaining state between the wireless stations and their interactions with Access Points is required for Intrusion Detection to be effective. The three basic states for the 802.11 model are idle, authentication, and association. In the idle state, the wireless station has either not attempted authentication or has disconnected or disassociated. In the authentication state, the wireless station attempts to authenticate to the AP or in mutual authentication models such as the Cisco LEAP implementation, the wireless station also authenticates the AP. The final state is the association state, where the wireless station makes the connection to the network via the AP. Following is an example of the process of maintaining state for a wireless station:

1. A sensor in promiscuous mode detects a wireless station trying to authenticate with an AP
2. A state-machine logs the wireless stations MAC address, wireless card vendor and AP the wireless station is trying to associate to by reading 802.11b frames, stripping headers and populating a data structure usually stored in a database
3. A state-machine logs the wireless station's successful association to the AP

State Analysis looks at the behavioral patterns of the wireless station and determines whether the activity deviates from the normal state behavior. For example, if the wireless station was broadcasting disassociate messages, that behavior would violate the 802.11 state model and should generate an alarm.

2.9 Multi-Dimensional Intrusion Detection

The very natures of Wireless LANs intrinsically have more vulnerabilities than their wired counterparts. Standard wire-line intrusion detection techniques are not sufficient to protect the network. The 802.11b protocol itself is vulnerable to attack. A multi-dimensional approach is required because no single technique can detect all intrusions that can occur on a wireless LAN. A

successful multi-dimensional intrusion detection approach integrates multiple intrusion detection models that combine quantitative and statistical measurements specific to the OSI Layer 1 and 2 as well as policy deviation and performance thresholds. Quantitative techniques include signature recognition and policy deviation. Signature recognition interrogates packets to find pattern matches in a signature database similar to anti-virus software. Policies are set to define acceptable thresholds of network operation and performance. For example, policy deviation analysis would generate an alarm due to an improper setting in a deployed Access Point. Attacks that exploit WLAN protocols require protocol analysis to ensure the protocols used in WLANs have not been compromised. And finally, statistical anomaly analysis can detect patterns of behavior that deviate from the norm.

2.10 Signature Detection

A signature detection or recognition engine analyzes traffic to find pattern matches manually against signatures stored in a database or automatically by learning based on traffic pattern analysis. Manual signature detection works on the same model as most virus protection systems where the signature database is updated automatically as new signatures are discovered. Automatic signature learning systems require extensive logging of complex network activity and historic data mining and can impact performance.

For wireless LANs, pattern signatures must include 802.11 protocol specific attacks. To be effective against these attacks, the signature detection engine must be able to process frames in the airwaves before they are on the wire.

2.11 Policy Deviation

Security policies define acceptable network activity and performance thresholds. A policy deviation engine generates alarms when these pre-set policy or performance thresholds are violated and aids in wireless LAN management. For example, a constant problem for security and network administrators are rogue Access Points. With the ability for employees to purchase and deploy wireless LAN hardware, it is difficult to know when and where they have been deployed unless you manually survey the site with a wireless sniffer or scanner.

Policy deviation engines should be able to alarm as soon as a rogue access point has been deployed. To be effective for a wireless LAN, a policy deviation engine requires access to wireless frame data from the airwaves.

2.12 Protocol Analysis

Protocol analysis monitors the 802.11 MAC protocols for deviations from the standards. Real-time monitoring and historical trending provide intrusion detection and network troubleshooting. Session hijacking and DoS attacks are examples of a protocol attack. Maintaining state is crucial to detecting attacks that break the protocol spec.

3. Conclusion

The importance of security in a wireless environment can not be under stated. Because the transport medium is shared, potentially beyond the physical security control of the organization permits attackers easy and unconstrained access. As a result, strong access control and authentication become essential in protecting the organization's information resources. IEEE 802.11 standard is updated day by day and the security aspects of wireless LAN is getting more importance than before.

Reference

- [1] LAN MAN Standards Committee of the IEEE Computer Society, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ANSI/IEEE Std 802.11, 1999.
- [2] Joshua Hill, *An Analysis of the RADIUS Authentication Protocol*, 2001.
- [3] W.A. Arbaugh, N. Snakar, and J. Wang. "Your 802.11 Network has no cloths, Proceedings of the First IEEE Intl. Conference on Wireless LANs and Home network, December 2001.
- [4] Arunesh Mishra, W. A. Arbaugh, *Security Analysis of the IEEE 802.1x Standard*, Proceeding, February 2002.
- [5] www.cisco.com



Some aspects of coordination paradigm in concurrent programming

D N Ranasinghe
Department of Computation and Intelligent Systems,
University of Colombo School of Computing,
Colombo, Sri Lanka
dnr@ucsc.cmb.ac.lk

Abstract

The separation of computation from coordination has been advised for a long time in the domain of concurrent programming. It has been shown to exhibit better reasoning and scalability in software design. In this paper we review the fundamental coordination models and their derivability from one another. Some encodings are already established in the literature and for others we postulate encodings based on logical inferences. The latter have not yet been theoretically proven.

Keywords: coordination models, abstractions, concurrent programming, software design, parallel computing, agent systems

1. Introduction

In concurrent and parallel computation, coordination refers to abstracting of higher level configurations away from computation thereby working towards better reasoning and higher scalability. In its narrowest form, coordination can be thought of as a form of collaborated computation by software components. Concurrent programming through coordination [1] has been studied by many [2,5,8,9,13,18,25,26,29] during the past years.

A coordination model is defined by a triple: E, the coordinated entities (agents, tasks etc), L, the media used to coordinate (shared variables, shared data spaces, channels etc) and M, the semantic framework the model adheres to (such as associative matching, set rewriting, logic laws etc) [25]. A wide variety of languages and models adhering to the coordination paradigm exist, and in this paper we intend to observe a particular property of these models: the derivability or the encoding of one model from another through abstraction. Theoretical work in this regard has been done on few of the models [14,20,22,24], but here, we use logical inference to argue our case for the derivability of the rest of the models.

In section 2, we present a classification of coordination models. In section 3, the fundamental models are briefly reviewed. In section 4, model derivability is investigated. In section 5, related work on model analysis is briefly reviewed. Section 6 concludes the paper.

2. A classification of coordination models

The concept of time-space coupling (or uncoupling) between a client process and a server process is central to the understanding of the categorisation of the possible modalities for interaction. In [33] four such categories are identified: ‘direct’ to designate time and space coupling (as in typical client-server computing), ‘mailbox’ for time uncoupled but space coupled (as in standard mail driven applications), ‘meeting oriented’ for time coupled but space uncoupled (as in event driven ‘TIBCO’ like systems) and finally ‘generative communications’ to indicate both time and space decoupling, to which a majority (but not all) of coordination languages belong. The key to time-space decoupling is the availability of a ‘persistent’ data space. This could be provided by a virtual shared memory with special properties on distributed-memory architecture.

In [25,26], the authors provide an extensive survey of coordination models and languages, in which two distinct and broad categories of coordination models are identified: data driven and control driven, where each category can be of either message passing or shared data space flavour. In the data driven category, concurrent computational threads are synchronized on the availability of data elements in the persistent data space, whereas in the control driven architecture, there is a separate configuration thread that manages the dynamic port connectivity between threads that possess ports for communication. Control driven models on follow Hoare’s CSP like architecture with interprocess communications through ports associated with a process [25]. Orthogonality

in computation and coordination has been cited as the main distinguishing feature between the two categories data driven and control driven.

As such there are four combinations of coordination models: data driven with message passing (e.g., Gul Agha's Actors), data driven in shared data space (e.g., Linda []), control driven with message passing (e.g., IWIM []) and control driven in shared data space (e.g., Structured Gamma [31]).

Models that run on the shared data space principle can work in one of three broad paradigms: associative matching set inclusion and set rewriting [17] which are not in fact mutually exclusive as we shall see later.

3. Fundamental models

3.1 Generative communication model

Linda [2] is the first known language of this category. It is based on a shared data space abstraction. From a practical point of view, we could imagine a collection of processes interacting with each other through a shared data space (also called a tuple space) while executing on a distributed memory architecture. [33] also sees this as one of the three (among page based, shared variable based and object based) approaches in distributed shared memory architectures (DSM). In this scenario, the processes are time and space uncoupled, in that they need not exist simultaneously in time nor know each other's locations. From a mathematical point of view, operations take place over a multiset, i.e., a set containing replicated elements, of tuples (collections of typed elements). There are three possible forms of multiset operations, on each of which a coordination model has been derived. We briefly discuss them now.

3.1.1 Associative matching.

Linda and its extensions such as Java spaces (a JINI service) and IBM Tspaces use associative matching where the typed fields of a tuple and its template are matched (in case of a 'read' primitive) to couple the interacting tasks. The tasks themselves are programmed in a host language such as C, Java etc where primitives of the type 'in', 'rd', 'out' and 'eval' are embedded within it. The following example illustrates this mechanism. Suppose there is a pool of integers out of which the maximum is to be found. The following code fragment when spawned as concurrent tasks will accomplish this.

```
Lindamax =    in (x1);
               while(true)
               {inp(x2);
                if (x 1 >= x2) then
                out(x1)else
                out(x2);
                in(x1);}
               return (x1);
```

In this model, coordination is not fully orthogonal to computation. In other variations we can find centralised and distributed tuple space implementations, and if distributed, they are to be with or without replication and partitioning.

Java spaces is a centralised implementation, but its primitives can support multiple tuplespaces through atomic transaction semantics. A particular feature of Java spaces is that its objects implementing the command pattern enables generic clients to run methods of tuple objects, even if the fetched objects are in a different language.

3.1.2 Set rewriting.

The computation models Gamma and CHAM (chemical abstraction machine) [17] are based on the multiset rewriting abstraction. In Gamma, a task consists of a pair of rules: a condition and an action. If the multiset is a subset of the condition then the multiset is rewritten according to the action. In this concurrent model, the computation is likened to a chemical reaction that takes place in a solution of chemicals. The previous example for finding the maximum can be simply coded in one line using Gamma:

```
Gammamax = { | x1, x2 ;
x1>= x2 | } | => { | x1 | }
```

3.1.3 Set inclusion.

This is a slightly obscure model and in the literature Bauhaus Linda [8] is mentioned as based on this approach. In this instance the shared data space consists of nested multiple tuple spaces, giving a hierarchical structure to the data space. The structure also leads to navigability in the space. In Bauhaus, tasks can insert and remove subspaces containing the desired tuple while locally consuming the same for computation purposes. An example of a Bauhaus operation is given later.

3.2 Message passing with configurations

This is also known as control driven coordination [25]. In this, a configuration code determines the dynamic connectivity between concurrent computation tasks, where the latter read from and write to pre designated ports. The metaphor is based on Hoare's communicating sequential process (CSP) abstraction. In fact this is an extension of the raw message passing paradigm but with space uncoupling. A notable fact here is the orthogonality of computation and coordination. As such whereas the configuration code has to be written in a specific language (which permits port adjustability), the computations can be written in any of the well known languages.

The following extracted example[25] shows the utilisation in a patient monitoring system.

```

group module patient;
  use monmsg: bedtype,
  alarmstype;
  exitport alarm:
  alarmstype;
  entryport bed: signal
  type reply bedtype;
  << computation code>>
end.

group module nurse;
  use monmsg: bedtype,
  alarmstype;
  entryport
  alarm[1..maxmsg]: alarmstype;
  exitport bed [1..maxbed]:
  signal type reply bedtype;
  << computation code>>
end.

system ward;
  create
    bed1: patient at
  machine1;
    nurse: nurse at
  machine2;
  link
    bed1: alarm to
  nurse.alarm[1];
    nurse.bed[1] to
  bed[1].bed;
end.

```

4. Observations on model derivability through examples

We take pair-wise combinations of fundamental models and look into the derivability of one in another. As mentioned before, many of these encodings have been formally established but, some have not been so. We use simple examples to check the ability to encode, but it has to be accepted that this is not a fully satisfactory method and more rigorous methods are advocated.

4.1 Known encodings in the literature

4.1.1 Linda in CHAM/ Gamma :

This has been widely studied. As reported by [14, 15, 24], provided Linda does not have a primitive that requires a global test for presence or absence of data, then this encoding is possible. The primitives which require such a test are *inp*, *rdp* and *notify*.

Eg [20]:

```

Suppose
  p:
  eval (q) .rd(a) .out(b) .end
  q := out(a) .in(b) .end

Then, CHAM(p) = { |
  eval (q) .rd(a) .out(b) .end | }
  = { |
    out(a) .in(b) .end,
    rd(a) .out(b) .end | }
    = { |
      in(b) .end, rd(a) .out(b) .end | }
      = { | a,
        in(b) .end, rd(a) .out(b) .end | }
        = { | a, in
          (b) .end, b, end | }
          = { | a, end,
            end | }
            = { | a, end,
              end | }
              = { | a | }.

```

The Linda calculus used for the above has been provided in [20].

4.1.2 Channels (control driven model) in Linda:

In its simplest form, a channel is a connection between two ports each belonging to a process. A process reads from or writes to a port. Here we take the IWIM-Linda [9] example. In this abstraction, with a single Linda data space, a channel is modelled as a forwarding process.

Eg. [9]:

```

Suppose a channel is defined as,
producer.outport -> consumer.inport
Then, forwarder (prod, cons,
outport, inport) {
    While(true) {
        in (prod,
outport, data);
        out (cons,
inport, data);
    }
}

```

whereas, the producer and the consumer processes would correspondingly do matching out (*or* in) primitives.

4.1.3 Channels in Bauhaus Linda:

In this abstraction, a channel is modelled as a tuple space. [9] mentions that Bauhaus is powerful enough to express the notion of a channel. Here we consider a simple example to demonstrate this fact.

Eg:

Let process P1 communicates with P2 over a channel named 'x'. In Bauhaus this can be expressed in the two primitives, P1: out {x -> a} and P2: rd {x} where they interact over the multiset {P1 P2 {x}} or over the multiset {P1{P2{x}}}.

Suppose a configuration process C wishes to establish a channel 'x' which does not initially exist. This can be described in Bauhaus as:

```

{ C {P2} {P1{x}}} -----> C: in
{P1}; out {P2 ->P1} ----->
{C{P2{P1{x}}}}

```

Here, the last multiset configuration enables P1 and P2 to communicate over channel 'x'.

4.2 Inferred Encodings

The following possible encodings have not been mentioned in the literature. We speculate as to the existence of them based on observations made earlier.

4.2.1 Channels in CHAM/Gamma:

This hypothesis can be proven correct by the existence of the following fact. With the observation that there exists a channel encoding in Linda and, a Linda encoding in CHAM/Gamma (section 4.1)

4.2.2 Bauhaus Linda in CHAM/Gamma:

This hypothesis can be proven correct by the existence of the following fact. With the

observation that there exists a channel encoding in Bauhaus and a channel encoding in CHAM/Gamma.

4.2.3 Bauhaus Linda in Linda:

Provided our above hypotheses are correct and in the absence of any proof to the contrary, we should be able to say that an encoding of Bauhaus Linda in Linda is possible. This is because, we already have an encoding of Bauhaus in CHAM and an encoding of Linda in CHAM (subject to certain conditions). But, apparently this does not seem to be very correct, as Bauhaus is known to be highly expressive than simple Linda.

We speculate that this hypothesis may be proven if we replace simple Linda with a Linda with first class properties. However this remains to be proven rigorously.

5. Related theoretical work on expressiveness of languages

In [25] Linda and Javaspaces which are based on associative matching have been formally expressed using Milner's CCS like calculus and SOS rules with the aim of identifying how each of the primitives and their variants in a language (such as notify and blocking with timeouts) can alter the expressiveness. In particular it has been shown that non blocking Linda primitives such as inp and rdp, which requires an examination of the global states removes the so called 'monotonicity' of the language affecting its encoding in another process calculus. It also shows that how notify encoded with a 'test for absence' with increase the expressiveness of Linda like languages [15].

In [20] a Linda calculus has been studied using SOS, CCS, Petrinets and CHAM, all of which are formalisms for modelling concurrency. [20] points out that CCS does not provide a realistic implementation of Linda due to the latter's associativity that creates unrestricted synchronisation whereas in CCS it is always a point to point synchronisation. As such higher order process calculi may be better suited for modelling Linda [20] suggests. The CHAM abstraction of Linda [20] can be considered as the first step in presenting an encoding of associative matching in set rewriting. The conclusion being that Linda is a form of multiset rewriting [20].

A formalism known as embeddings has been used to compare several coordination models and their possible implementation architectures by [23]. Varied architectures involve centralised, distributed, replicated and broadcast updates mechanisms and are

classified into locally delayed, globally delayed and undelayed implementations of primitives. Comparisons are made between a shared dataspace model, a so called distributed dataspace (sets with local consumption) and a control driven model [23]. They show that, for any given implementation architecture, the data driven models (in the absence of nonblocking primitives) and the control driven model are observationally equivalent. They also show that, non blocking primitives can only be implemented in a particular architecture [23].

The encoding of Linda in Gamma and the encoding of Gamma in Linda has been formally analysed by [15]. While it concludes that an encoding of Gamma in Linda is impossible as multiset rewriting rules are automatically executed in Gamma (and not so in Linda), the encoding of Linda in Gamma may be possible if the monotonicity property is kept intact as reported elsewhere. This requires that primitives with tests for absence of data are not implemented in Linda [15].

In [24] an exhaustive comparison of expressiveness of three shared dataspace models (all data driven) has been conducted using ‘modular’ embeddings, a form of language encoding. Linda, Gamma and a model containing ‘sequences of primitives with atomic transactions’ have been considered each with three types of variants of primitives. The variants signify those with and without a test for presence or absence of data. Results [24] corroborate previously obtained observations [15] that, Linda without a global check for absence of data is encodable in Gamma and, also less expressive than Gamma. They also show that the introduction of transactions to atomically execute a series of primitives increases the expressiveness of language. This is an agreement with the earlier outcome that it was the absence of atomic transactions in Linda which made Gamma to be encoded in Linda. It is suggested [24] that Polis [16] falls into this higher expressive category by virtue of the fact that it is based on Gamma like reactions (in hierarchical dataspace), it is also likely that reactive spaces [11] which execute reactions atomically will exhibit a similar higher expressiveness.

Deviating from the study of expressiveness [19] defines an abstract Linda system (ALS) to study the extent of conformance of many Linda like languages to ‘real’ Linda. This is done by expressing any Linda like system as a reactive system whose externally visible actions correspond to ALS actions.

Multiple tuple spaces constitute an important extension of functionality of shared dataspace models as shown by Bauhaus Linda [8] and KLAIM [13]. Bauhaus employs multiset inclusion in a hierarchical tuple space, whereas KLAIM uses Linda primitives extended with names in a flat multiple tuple space. Polis [16] extends the Gamma like reactions in a single space with nested multiple spaces where reactions takes place within a defined scope (similar to membranes in CHAM [17]. Multiple spaces could also be augmented with first class properties, where primitives could operate on spaces as first class tuples as shown by KLAIM [13] and [27]. This is expected to increase the expressiveness of the language, but no formal analysis exist in this regard. In [22] a calculus for a named flat multiple tuple space is given, without the important ‘new’ (creating a space) primitive. KLAIM [13] consists of a complete π like calculus for the multiple tuple space with first class properties. KLAIM is mainly targetted to model mobility and can be considered as a calculus for the same. Nothing has been reported about its ability to encode other coordination languages.

Reactive spaces [12] which combine two fundamental dataspace paradigms, and its various language implementations have been targetted for mobile agent interactions. However, apart from a test for Turing completeness of the language based on reactive spaces, no further formal analysis has been done.

6. Final remarks

This paper has presented a brief introduction to the coordination paradigm in concurrent programming. In doing so it has classified the domain of coordination models and has investigated the derivability of one model in another. Past efforts have been made in providing language encodings among fundamental models. We have made an attempt to reorganise this effort, thus identifying instances of encodings which have not yet been theoretically proven, but may exist hypothetically. As for the applications of coordination models there are many. In parallel computing, skeletons have been found a useful paradigm. These skeletons have been written as coordination and computational parts using a functional metaphor. In multi agent interaction, the time and space uncoupled property has been utilised to maximum to provide flexibility. Distributed simulation has been performed using control driven models. Control driven models are specifically targeted for large scale software design and modelling.

Recently, web and peer to peer based extensions of the dataspace concept have been proposed as well.

References

1. Models of coordination; R. Tolksdorf; LNAI 1972; 2000
2. Coordination languages and their significance; Communications of the ACM; 35(2); D. Gerlernter, N. Carriero; 1992
3. Logic channels: a coordination approach to distributed programming; M.Diaz, B.Rubio, J.M Troya; IPPS 97; 1997
4. Using logical operators as an extended coordination mechanism in Linda; J.Snyder, R.Menezes; Coordination 2002; LNCS 2315; 2002
5. The shape of shade: a coordination system; S.Castellini, P.Ciancarini, D.Rossi; Technical Report UBLCIS-96-5; Univesrsity of Bologna; 1996
6. Interaction abstraction machines; J-M Andreoli, P. Ciancarini, R.Pareschi; Research directions in concurrent object oriented programming; MIT Press; 1993
7. Programming by multiset transformation; Communications of the ACM; 36(1); J-P Banatre, D.Le Metayer; jan 1993
8. Bauhaus Linda; N. Carriero, D. Gerlernter, L.Zuck; Object based models and languages for concurrent systems; LNCS 924; 1995
9. Control driven coordination programming in shared dataspace; G.A.Papadopoulos, F.Arbab; Proc 4th Intl. conference on parallel computing technologies; 1997
10. Tuple based technologies for coordination; D. Rossi, G.Cabri, E.Denti; Coordination models and applications for agents; Springer; 2001
11. Programmable coordination media; E.Denti, A.Natali, A.Omicini; Coordination 1997; LNCS 1282; 1997
12. On the expressive power of a language for programming coordination media; ACM SAC 1998; E.Denti, A.Natali, A.Omicini
13. KLAIM; a kernel language for agent interaction and mobility; R.de Nicola, G.L.Ferrari, R.Pugliese; IEEE transactions in Software Engineering; 24(5); May 1998
14. On the semantics of tuple space based coordination models; ACM SAC 1999; A.Omicini
15. On the incomparability of Gamma and Linda; G. Zavattaro; CWI technical report; SEN-R9827; 1998
16. Using a coordination language to specify and analyse systems containing mobile components; P.Ciancarini, F.Franze, C.Mascolo, ACM transactions on Software Engineering and Methodology(2000)
17. Gamma and the chemical reaction model: 15 years after; J-P Banatre, P.Fradet, D Le Metayer; Multiset processing; LNCS 2235; 2001
18. Distributed coordination with messengers; M.Fukuda, L.F.Bic, M.B.Dillencourt, F.Merchant; Science of Computer Programming; 1997
19. On what Linda is: Formal description of Linda as a reactive system; D.Gerlenter, L.Zuck; Coordination 1997; LNCS 1282
20. On the operational semantics of a coordination language; P.Ciancarini, K.K Jensen, D.Yankelevich; In object based models and languages for concurrent systems; LNCS 924; 1995
21. Software architecture styles as graph grammars; D.le Metayer; ACM SIGSOFT 1996
22. Models for coordinating agents: a guided tour; N.Busi, P.Ciancarini, G.Zavattaro; Monograph on coordination of internet agents, models, technologies and applications; Springer; 2001
23. Comparing coordination models and architectures using embeddings; N.M Bonsangue, J.N Kok, G.Zavattaro; CWI Report SEN-R0025; 2000
24. On the expressiveness of coordination models; A Brogi, J-M Jacquet; COORDINATION 99; LNCS 1594; 1999
25. Coordination models and languages; G.A.Papadopoulos, F Arbab; CWI technical report SEN-R9834; 1998
26. Coordination models and languages; In objective coordination in MAS engineering; LNAI 2039; M Schumacher (Ed); 2001
27. Customisation of first class tuple spaces in a higher order language; S. Jeganathan; PARLE 1991, vol2; LNCS 506
28. Coordination with scopes; I.Merrick, A.Wood; ACM SAC 2000
29. Actorspaces: an open distributed programming paradigm; G.Agha, C.J Callsen; SIGPLAN notices; 28(7); 1993
30. Mixed programming metaphors in a shared dataspace model of concurrency; G-C Roman, H.C Cunningham; IEEE transactions on software engineering; 16(12); 1990
31. Structured Gamma; P. Fradet, D. le Metayer; Science of Computer programming; 31(2); 1998
32. Implementing skeletons for generative communication with Linda; D.K.G Campbell; University of York technical report, UK; 1997
33. Distributed Systems: principles and paradigms, Andrew Tanenbaum, Prentice Hall, 2002



An Approach to eTransform Enterprises in Developing Countries (A Case Study of an Enterprise in the Ceramics Sector in Sri Lanka)

Mahesha Kapurubandara¹

Shiromi Arunatileka²

Prof. Athula Ginige³

University of Western Sydney, Locked Bag 1797
Penrith South DC NSW 1797, Australia

¹Email : mahesha@cit.uws.edu.au

²Email : shiromi@cit.uws.edu.au

³Email : a.ginige@uws.edu.au

Abstract

Developing countries differ from their affluent counterparts, the “developed”, in numerous ways. Infrastructure, cultural, social and regulatory differences are among the main factors. These differences or barriers tend to widen the digital divide. They stand in the way of the developing countries trying to achieve their goals towards a global economy by embracing eTechnologies. The feeble and many unsuccessful attempts to re-cycle the methodologies used by the developed countries, have left the developing high and dry. In formulating strategies for e-transformation of developing countries, the barriers specific to countries with lower GDPs have to be taken into serious consideration.

In this paper, an eTransformation model that is being successfully used with SMEs in Australia is being modified appropriately, proposed and applied as the approach for eTransformation for developing countries using a case study approach. The 7E's in eTransformation is a model developed by researchers at the University of Western Sydney. It is currently being used successfully with a group of companies in Western Sydney. The model incorporates new business thinking, business models in the new e-economy and addresses issues such as analysing the external environment in eTransforming, re-engineering business, business-IT alignment, and change management issues. A company in the ceramic manufacturing sector in Sri Lanka – is being used as the case study for eTransformation.

Key words: e-transformation, e-business, developing countries, 7Es in eTransformation

1. Introduction

The Internet has entered every sphere of human life. Business is included. E-Business and E-Commerce have revolutionized ‘buying’ and ‘selling’ throughout the world. Competition is now not among individuals but among nations. Information Communication Technologies (ICT) and the economy have become so heavily dependent on each other that it has become very necessary to re-evaluate our business and economic environment. Big and small, developed and developing, every nation faces stiff competition in the World Market, which is now a Virtual Market. The Virtual World, with its numbers of host companies offers serious challenges to SMEs (Small and Medium scale Enterprises) all over. Never before have businesses, especially SMEs faced such heavy open competition both global and internal. They are now driven to meet challenging demands from both customers and business partners.

The concept of a global village has revolutionised the traditional business scenario. To compete in this atmosphere and look for profit, success and sustainability, the use of Information Communication Technologies is imperative. ICT has risen to the fore as an effective strategic business tool to gain competitive advantage and remain solvent.

2. Benefits of eTransformation to Developing Nations

In order to exploit the global online market, there is no need to necessarily come up with a mega solution such as Amazon.com or eBay. Different business models and strategies combined with new thinking can do wonders to provide on-line services to millions of customers globally. Some benefits of eTransformation for developing countries are: business growth in the

global market, building strategic alliances/partnerships, economic growth by revenue generation, strengthening the SMEs, infrastructure development, employment generation, improving accessibility to the world market, etc. In an essence, eTransformation has the trickle down effect, which could be seen running across industries, ultimately bringing improvement to the quality of life of the people.

3. Problems faced by Developing Countries

The Net is capable of generating higher incomes and higher standards of living. However, multinational companies invest only where communications infrastructure is reliable. As a result the digital divide keeps growing, driving the less fortunate even further from their goals. Already this divide includes a larger part of the population, especially in developing countries. Therefore, E-Commerce may become a trade barrier for those not connected.

The following statement made by the UN ICT Task Force, Geneva node, can be used more aptly to describe at the root level, the digital divide between the developing and developed countries with regard to economic and social betterment.

“Today, less than 15 years after the fall of the Berlin Wall, a new divide is appearing in Europe, the divide between those who have access to information and communication technology and those who have not. This 'Digital Wall' is beginning to separate countries, regions, cities and people in terms of economic and social development.”

Furthermore, in developing countries a majority of business happens to be SMEs. They dominate the economy of the country. There is huge potential for these enterprises to grow and shown the proper way, reach the global market. They can sell their products in international markets; become sub contractors to the market giants etc. The opportunities are unlimited, if they are shown the way to tap their resources.

A survey of SMEs in the APEC region finds legal and liability concerns ranking right behind a lack of market demand and security concerns. These issues are considered the topmost barriers to the development of e-commerce and are seen more among the lower-GDP economies as against the higher GDP economies[4]. To eradicate these barriers it is necessary to improve telecommunication infrastructure, reduce legal barriers, and increase business access to the Internet in order to boost the use of e-commerce by SMEs[4].

Against this general background, it is not surprising that many enterprises in developing countries are still sceptical about the wider use of ICT

and e-business. It is observed that SMEs need a better understanding of the opportunities e-commerce provides for their businesses. The cost of accessing the e-market is high and one is not assured of benefits.

All these arguments contribute to explain why eTransformation of SMEs in developing countries is not simply re-cycling strategies, methodologies and products that work in the affluent, developed world. New models and methodologies have to be designed for the

E-Transformation of SMEs in the developing regions keeping in mind the barriers and constraints mentioned above. Some barriers for eCommerce development among the developed and developing are illustrated in the following table.

Developed Countries	Developing Countries
<ul style="list-style-type: none"> • Infrastructure- Reliable, adequate (Electricity, Telecomm., ICT) Cheap internet, Easy access to phones • Economy Financially stable • Social & Cultural High digital literacy English used as the medium for global business • Business culture- Virtual trading Highly developed related and supporting industries • Regulatory Proper government industrial policies, internet policies, laws and legislations, national information policies 	<p>Inadequate Infrastructure</p> <p>Unstable, lacking financial resources</p> <p>Relatively low or poor literacy, English language barrier</p> <p>Face to face trading Relatively poor development</p> <p>Policies non-existent or not adequate</p>

Table 1 Comparison of Social and Cultural Differences between Developed and Developing Countries - Barriers for e-commerce Development

The table illustrates the differences pertaining to developing countries in general. The resources, facilities, and other factors which are taken for granted in developed countries, are either lacking or do not exist in developing countries.

4. A Successful Approach to eTransformation

An extensive study of existing eTransformation methodologies was carried out by the researchers at the University of Western Sydney, Australia[6]. Due to lack of a complete model which looks at all important

aspects of eTransforming, a model, namely, Seven E's in eTransformation[6] was developed. The 7E model has been applied successfully to transform enterprises, in Western Sydney, Australia,[13] in the context of developed countries.

This paper makes an attempt to modify the 7E model, which has been successfully tested with enterprises in developed countries to suit the conditions in developing countries. To test the approach, Sri Lanka- a developing country in Asia is chosen. It is opportune as Sri Lanka is trying to re-gain itself and there is tremendous interest shown towards IT and ICT. This study applies the 7E model to an enterprise in the ceramic manufacturing industry in Sri Lanka.

The environment the SMEs work is changing constantly. Analysing the environment and the global IT/business trends are crucial for strategic e-transformation to find out the best opportunities for the SMEs to invest in.

The Seven E model commences the process by analysing the external environment and deciding on broader eBusiness goals and strategies for the selected sector. Priority is also given to checking the e-readiness of the sector before proceeding with the transformation. The Roadmap for eTransformation and a methodology successfully used with the SMEs in Western Sydney are used as the vehicle for eTransformation of the selected enterprise dealing with the ceramic industry in Sri Lanka. High priority is also given to eSystems, support services and change management issues making sure that the transformation is carried out successfully.

5. The Seven Es In eTransformation and its Applicability to Developing Countries

This model consists of seven very important aspects of eTransformation. Each stage is important in its own right and forms a part of the whole process. The seven stages, where six stages could be achieved one after the other are linked to the stage 'Evolution'. This deals with the crucial issues related to change management (Figure 1). After each stage, the organization can go through the changes to the evolution stage and through that, go to the next stage, after the required evolutionary changes are made.

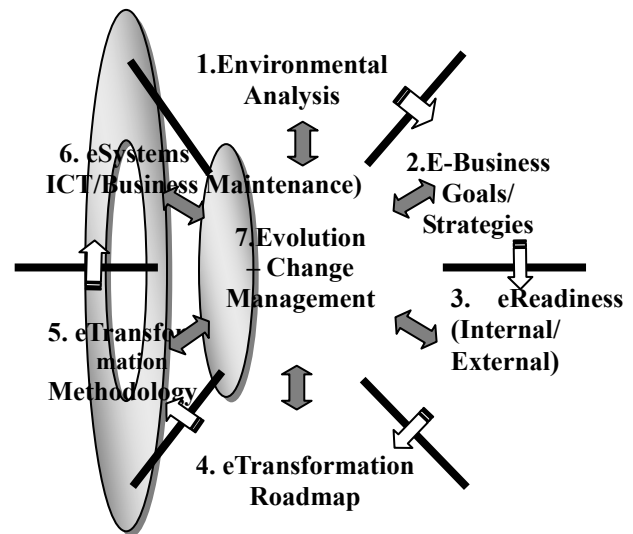


Figure 1. The Seven Es in E- Transformation [6]

Since 1999, the researchers at the University of Western Sydney has been conducting a research project in eTransforming Western Sydney. Since the Seven E model is being used in a different context, for enterprises in a developing country, the need for modifying the methodologies used within each stage is looked at carefully. The methodologies used within each stage is briefly illustrated below (Table 2).

Stage	Significance	Methodology used	Strengths	Weaknesses
1. Environmental Analysis	To understand the Global IT and Business Trends and the Sector's Strategic Situation	Understanding of Global Trends in Business and IT SWOT Analysis Industry Analysis (Porter's Forces)	* Cheap abundant labour * Developed countries' willingness to invest on ICT in developing countries * Strong manufacturing sectors	* Infrastructure limitations * Telecommunications and Access * Legal framework for eCommerce
2. eBusiness goals and Strategies	Develop eBusiness goals and strategies to gain competitive advantage	Develop eBusiness, Strategies & Adopt eBusiness Models	* Many businesses having a stable buyer market in developed countries * Educated literate management workforce	* Awareness in eBusiness in industries * Funding needs * Digital divide (within/ across country)
3. eReadiness	E-Readiness of the industry and the enterprise under consideration	Questionnaires to measure eReadiness of the Internal and External entities	Opportunities * No geographical boundaries in eMarketplace * Apply new e-business models - bundling/ unbundling of services * Direct-to-customer/ market approach * Exporting IT-based services * E-Portals for SMEs in manufacturing * Income/ Employment generation * Incubators/Tele-centres * E-Learning Opportunities * Attracting ICT products/ services based investments from industry giants * Improve accessibility * Overall economic growth * Improve Infrastructure * Open-source software giving a less costly start to offering IT services * Access to global economy	Threats * Difficulty in competing in the already established e-marketplaces * Language barrier to provide services * The need to be physically close the real market (for shipping) * Speedy development of ICT/services * Telecom. access within countries creating bottlenecks * Political problems in developing countries limiting growth * Convincing the buyers of the quality of products/services * Ineffective implementation of Government Policies/regulations
4. eTransformation Roadmap	To develop a specific path to proceed for the organisation	eTransformation Roadmap and the Convergence Model		
5. eTrans. Methodology	eTransform the organization in an incremental way	The Evolutionary eTransformation methodology		
6. E-Systems	Provide support and maintenance of the implemented systems	Develop IT Policies, Security, Support, Maintenance mechanisms		
7. Evolution – Change Mngmnt.	Management of the proposed changes in an evolutionary manner	McKensy's 7S Model for Organisational Change		

Table 2: Stages in the 7Es Model: Modified to suit Enterprises in Developing Countries

The model incorporates new business thinking, business models in the new eEconomy and addresses issues such as analysing the external environment in eTransforming and re-engineering business, Business-IT alignment, implementing and managing systems, restructuring, change management, e-systems maintenance and policy issues. The model can be used to successfully eTransform an organisation, a cluster of enterprises or an industry to achieve profitability in eBusiness.

5.1 eTransformation related Issues in Developing countries

First, we need to understand the different conditions under which developing countries operate. In order to compare how conditions differ from those of the developed countries we carried out an eBusiness related SWOT Analysis for developing Countries[Table 3].

Table 3: E-Business related SWOT Analysis for Developing Countries

The above SWOT analysis (Table 3) shows that there is a lot of potential for enterprises to benefit from eTransformation. It also shows the major barriers and threats faced by enterprises in developing countries. One outcome of the analysis is that it reflects that issues to be addressed in four levels. They are:

- Issues at Organisational Level
- Issues at an Industry Level
- Issues at a National Level
- Issues at a Global Level

Issues to be addressed at organisational level are internal transformation issues such as infrastructure, business processes, management support, staff and skills development, etc. At an industry level, the issues

are, formation of strategic partnerships, quality of products/services offered, infrastructure related issues, creation of awareness in eBusiness, etc. At a national level, the government support for infrastructure development, tax incentives and for the implementation of government policies and regulations are essential. At a global scale, the issues and concerns are for international funding, international partnerships for infrastructure development, usage of Online and offline marketing strategies and web based strategies [12].

5.2 The Applicability of the 7E Model to Enterprises in Developing countries

When applying the 7E Model, the conditions of the environment are automatically studied (in stage1) and incorporated in to the development of strategies (in stage2). It is important to study the external environment as well as the internal environment for each stage. The following section describes the modifications needed at each stage of the model:

Stage 1 : Study the external environment in a detailed manner eg: carry out a SWOT analysis for the industry or the country depending on the application. The 7E Model already does an extensive analysis at this stage which could be extended to look at the environment in a global context.

Stage2 : Development of strategies at a national level to deal with funding, infrastructure and other needs. It has to be taken as a national initiative more than a mere company transformation.

Stage 3: Emphasis should be given to the eReadiness of the country, industry and the external entities with which the company interacts with.

Stage 4: The roadmap is applicable to any organization or a cluster or organizations as it uses a step by step incremental approach to reach the eTransformed state.

Stage 5: The evolutionary nature of the methodology uses an incremental approach to eTransformation which suits the conditions of the enterprises in developing countries.

Stage 6: The systems in the original version of the 7E model mainly concentrates on the internal systems. When it is being applied to developing countries, the knowledge gained about the external environment in other stages have to be applied to deal with lack of infrastructure and support services.

Stage 7: The flexibility this stage gives to all other stages is the key to organisational success in dealing

with difficult environmental conditions. Depending on the changing internal/external conditions, the flexibility is given for companies to adapt to the changes in strategy, structure, systems, skills, staff, style and shared values.

In essence, the 7E model can be applied successfully to enterprises in developing countries with the model expanding to concentrate on the external environment at each stage for analysing and developing strategies. The application of the model to an enterprise in the Ceramics sector is described in the following section as a case study.

6. The Ceramic Industry in Sri Lanka

Porcelain production is a sector that flourishes in Sri Lanka. These products have carved a niche for themselves in the market with high quality at relatively low cost. A key factor to the success of the ceramic industry in Sri Lanka is due to the fact that about two-thirds of the raw materials needed are produced locally. With a vital cutting edge of impeccable quality, competitive pricing, unique creativity and a keen market driven sensitivity to modern trends and needs, Sri Lanka competes effectively with the world markets with ceramic front-runners such as Italy, Spain, China, and Indonesia. Key markets have opened up for Sri Lanka ceramics in USA, Australia, the Maldives.

The trend in ceramic industry is now for cheaper, colourful and casual tableware. This change of trend has affected the manufacturers. The product lifecycle for new shape and design has decreased and the manufacturers are forced to bring out new products on a regular basis in order to maintain the market share. The industry faces numerous constraints, including a lack of preparedness for use of advanced technology and expanded markets.

This paper uses a case study of an enterprise in the Ceramics industry using the 7E model to eTransform in order to gain the competitive advantage. The following section describes the application of the seven stages to the selected manufacturing organisation.

7. eTransforming the Selected Enterprise

We selected an enterprise in a developing country, Sri Lanka, to eTransform and the 7E Model is used as the model taking the company through the eTransformation process.

The company studied has been in the ceramic manufacturing business for more than a decade and now boasts about an employee strength of 600. It

exports 90% of their produce to 20 countries including France, Germany, Spain, USA, Italy, Portugal and Japan. It produces 500,000 pieces (equivalent to about 1,500 tonnes) of porcelain tableware every month.. The company is currently at peak capacity and also enjoys a growing market. Due to the high quality of the product and excellent customer service the demand for the product has grown to such an extent that the company cannot fulfil all orders that are received. As a result an expansion programme is now being planned.

7.1 Stage 1: The Environmental Analysis

The Sri Lankan Ceramics industry consists of about 40 ceramic export companies in operation including SMEs. They export ornamental ware, tableware, wall/floor tiles and sanitary ware. Sri Lanka's global market share is about 1%. This industry employs 22,000 individuals in the above segments of the industry [7].

In the UNCTAD report on E-Commerce and Development, it is stated that in the Asia and the Pacific regions, the manufacturing sector is exposed to pressure from customers in the developed countries to adopt eBusiness methods. It also emphasises on the value of eBusiness for intra-regional and global trade.

Michael Porter's Five Forces analysis (bargaining power of buyers, bargaining power of suppliers, threat of new entrants, threat of substitutes, rivalry among competitors) is being done in order to find out the competitive forces working on the industry. Depending on the forces, the adoption of strategies can be selected.

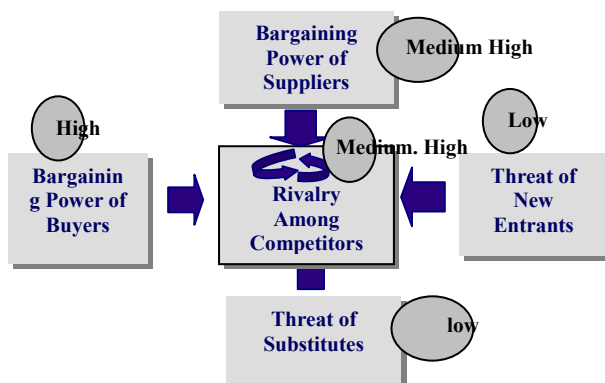


Figure 2: Applying Michael Porter's Five Forces Model to the Ceramic Industry

Michael Porter's Strategic Forces show that the highest threat is from the buyers as they keep looking for cheaper, colourful and casual tableware rather than the formal elegant tableware, which used to be popular. This change of trend has affected the manufacturers.

The product lifecycle for new shape, design has decreased and the manufacturers are forced to bring out new products on a regular basis in order to maintain the market share. Furthermore, the profit margins have also come down and all manufacturers are now looking at producing at the lowest possible cost. The other major force is the bargaining power of suppliers of fuel and the few expensive raw materials that have to be imported. This has a direct impact on the product life cycle, the production planning and the cost of the products.

Table 4 illustrates the SWOT analysis we carried out for the ceramics company.

Strengths	Weaknesses
Industry knowledge of CEO Manufacturing flexibility Company culture-best practices Innovation and creativity Customer base – Client pedigree Industry reputation Solution provider – design capability Low Fixed Costs Global customer relationship Good manufacturing technology	Web not used for competitive advantage Manual quality systems Lack of sales& marketing strategies IT is not used as a strategic tool Access to funding for growth Relative low wages Limited skilled staff
Opportunities	Threats
Increasing global market Closure of factories in the developed countries in the industry Access to modern technology Possibility of acquisition New product/market development Develop product range to a niche market Alliance with giants in the industry E-Business opportunities Expansion of customer base	Raw Material (Oil /Gas price) increases Market intelligence (Customer/competitor) Strong competition with cheap products Changing lifestyles No direct link to end-user No segmented marketing strategy

Table 4: The SWOT Analysis - for the ceramic company

7.2 Stage 2: eBusiness Goals and Strategies

The SWOT Analysis (Tables 3 & 4) gives the overall strategic situation of the companies in most developing countries in the Asian and Pacific regions. According to the research carried out by us, it shows that, in order to take advantage of the opportunities, the sector has to come up with eMarket strategies and link them to

eBusiness Models. For the whole industry to develop, we have to concentrate on funding, partnerships, infrastructure development, government support, legal framework, creating awareness, marketing strategies, etc.

eMarket Strategies: The industry sectors are changing the way they do business by using many different collaborations with customers (B2C), service providers (B2B), funding organizations (ePayments), government (B2G), and even competitors (B2B). The linear model of supplier-manufacturer-distributor-customer in the old economy is changing to adopt different e-business models [10] such as e-portals, Supply Chain Model, Full Service Provider Model, in the e-economy.

The information collected through interviews and questionnaires, reveals that, the goals of the ceramic company to e-transform are to increase the market share, to increase the quality of customer service, eliminate bottlenecks & reduce costs, to improve on supplier relationship management and to gain competitive advantage.

The best eBusiness strategies for this company after considering Porter's forces and the above goals are product differentiation, supply chain management, customer relationship management, marketing of products and strategic partnerships for joint ventures. The most suitable e-business model would be the Supply Chain eBusiness Model which creates a virtual value chain and an information flow across the supply chain. All parties have a strong electronic bond and backend systems and the manufacturers have access to information about the suppliers up to the level of the customers which is very effective in order processing, product tracking and SCM issues. It is a necessity to use effective Customer Relationship Management techniques to improve the quality of service to the customers.

- Supply chain E-Business Model (Horizontal Marketplace):

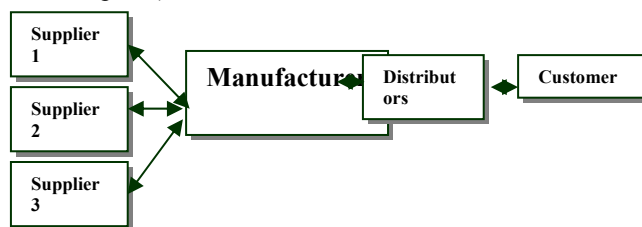


Figure 3: Supply Chain E-Business Model

Depending on the market dynamics and the competitive forces working on the industry, combinations of e-business models can be selected for implementation.

7.3 Stage 3: eReadiness

Seven important aspects need to be analysed in relation to the e-readiness of the organizations in the sector and the expected users of the web based system. They are Business Processes, Applications & Infrastructure, Web Presence, Skills, Top Management Commitment, External Connectivity and Future Directions. A questionnaire was given to the company. The results of the survey is presented below.

- Business Processes - Well Defined
- Applications & Infrastructure - Insufficient Resources
- Web presence - Presence does not serve the purpose
- Skills of Employees - Has a shortage of IT staff
- Top management commitment - Committed to productivity and excellence
- External connectivity - Main mode of communication is Fax and telephone
- Future directions - Directed towards usage of eBusiness Systems for buyer interactions, marketing, supplier related interactions and exploring business opportunities.

The support given by the ISPs (Internet Service Providers) along with their reliability and quality is also crucial for the successful implementation of the web-based systems. The bandwidth, web based services, dynamic content, ability to host databases, file transfer mechanisms, fees for hosting are some important aspects to be considered in making the decision to select the right service provider.

7.4 Stage 4: eTransformation Roadmap

After going through the first 3 stages, the company is fully aware of its strategic position, competitive advantage is and its readiness to e-transform. What it needs is a clear path to follow. The road map assesses the current status of the company and shows the direction to proceed.

The E-Transformation Roadmap [1] developed by the researchers at the University of Western Sydney, is the guideline used for successful e-transformation of many enterprises in the Western Sydney region (Figure 4).

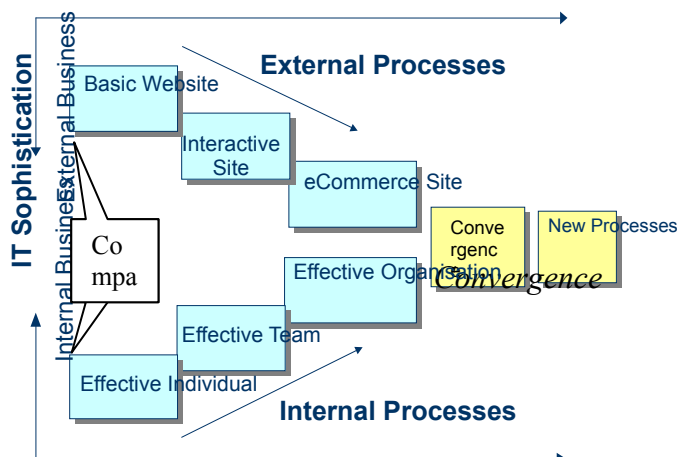


Figure 4 : E-Transformation Roadmap [1]

The final goal of eTransformation roadmap [figure 4] is to achieve the level of convergence. The convergence level is where all the information related to the business processes in an organisation (such as the information for marketing, sales, financial, production, etc.) are linked to a corporate data repository. The interactions with the external community such as the buyers, suppliers and other partners are through the integrated corporate data repository.

The ceramic company is currently in a very early stage of the eTransformation process. It has a very basic static website with the company profile. It would be advantageous for the company to start with a major improvement to the web based system expanding it to be used as a tool to gain competitiveness. This can be done in three phases.

Phase 1: Use of the Web presence as a Marketing Tool:

Building an attractive web based system and using the tested, proven online/offline methodologies could be used to effectively market the company and its products in the global market.

Phase 2: Include the Interactive Web Features:

To link with suppliers and/or customers depending on which link gives more benefits. Since many buyers are moving into newer designs, shorter production cycles and smaller quantities, the web can be used very effectively to give all the crucial information the buyers need to make quick and correct decisions.

The solution of this phase would be an eye-catching website with the following features: Company profile, Competitive advantage of the company, eCatalogue of the products with pictures, Site map/content information/search facilities,

Company policies, Customer specific information and security policies, Online ordering/shopping cart, News/FAQs, Forms, Production capabilities/customers, Special login for existing customers, etc.

Phase 3: Payments:

This will need to incorporate the ePayments systems thereby looking at security aspects and payment standards as well. In order to implement the above system successfully, the company has to think of internal process transformations such as incorporating the usage of a web based system, e-mail communications, effective team work using ICT.

7.5 Stage 5: eTransformation Methodology

At this stage, there are 2 paths to take, namely internal process transformation and/or external process transformation. The relevant approach for this company immediately is the external path as the company can effectively use the web based system to be used as a business tool to create awareness and to strengthen the relationships with the customers. The following methodology could be adopted.

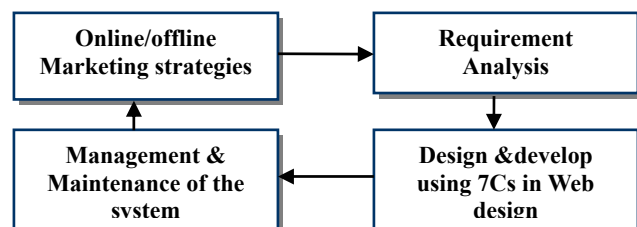


Figure 5: Adapted from the eTransformation Methodology [6]

The requirement analysis identifies the purpose of the website, the target audience, the need for interaction and its main functionality. The Design and Development stage identifies the web design aspect using the 7Cs in web design introduced by Rayport[9] It focuses on context, content, customisation, connectivity, communications, connection and commerce to deal with traditional customers, cyber customers and hybrid customers effectively. As these strategies are mainly for SMEs and SMEs do not have highly skilled IT staff to manage, maintain and update the web based systems, the designed system has to incorporate content management features as well.

Other aspects to be considered are updating frequency, automating to ensure responsibility, security and password protection, load balancing back up, etc. After all the arrangements are made for self-manageable websites, the marketing aspect has to

be looked into. Many online and offline marketing strategies such as business cards, including magazines, news letters, and search engines, banner advertisements, virtual marketing etc could be used to promote the web based system.

7.6 Stage 6: eSystems

After the Business Process Re-engineering, there will be a proposed organization-wide web based 'Business and IT' integrated system. They need to be supported by IT policies, management /operational controls & security measures, systems, etc.

Some management controls will incorporate standard guidelines to the users, procedures and manuals for the new system. Security measures are taken to deal with common threats such as sabotage, hacking, privacy problems, etc. Preparing for contingencies and disaster recovery are also done at this stage.

Since proposed strategies for the ceramic company would be the proper backup strategies, security measures, password protection, it has to be assured that the ISP or the computer vendors provide the proper trouble shooting and maintenance support to the company.

7.7 Stage 7: Evolution

This stage runs across all stages linking them to each other and implementing the Strategic, Managerial and Operational changes.

The model used to deal with the strategic, operational and other changes is the well accepted 7S model developed by McKinsey and Company over 20 years ago. The following issues were addressed with respect to the ceramic company. Shared values, Strategy, Systems, Structure, Skills, Staff and Style.

The ceramic company has to focus mainly on infrastructure development, structural changes and IT skills development of their staff. The best approach is to start staff training on Internet based basic skills such as web browsing, e-mailing, etc, and get the staff used to the idea of dealing with customers over the net.

It is very important for the enterprises in developing countries to focus on change management issues as they need to focus on the development of infrastructure, staff, skills, systems, funding and web based systems development more than SMEs in developed countries.

The issues addressed are as follows:

- Shared values: The Ceramic Company must ensure that the changes the company is going are

reflected properly in their value systems, their vision, mission and the organizational culture.

- Strategy: The strategies proposed were at the business level and the company has to break them down into management strategies and operational strategies for the company to successfully implement them.
- Systems: The systems are going to be changed tremendously due to the introduction of web based systems. The business process flow, the ordering systems, the storing systems, the response times, etc. will be changed and the thinking of the people and the business operations also need to change accordingly.
- Structure: The organization structure also needs to change from a rigid hierarchical structure to a more flexible network structure. It may be very difficult to change the structure, but, an effort needs to be taken to make the structure flexible enough to incorporate all the changes to the systems.
- Staff and Skills: This is a major area to concentrate on from the beginning as SMEs in developing countries lack IT staff and IT related skills as mentioned by the Ceramic company dealt with. They also require to train internal staff than to hire IT personnel externally as it is very costly for the company compared to their size.
- Style : IT is very important to have a non-authoritative management style as the company is going through a major change and needs the support of all levels of staff to work as a team to achieve a common goal of eTransforming the organisation.

8. Lessons Learnt

The application of the 7E model to the Ceramic manufacturing sector in Sri Lanka revealed a wealth of information. It reconfirmed the fact that eTransformation is a step by step evolutionary process for any enterprise depending on its e readiness. The study revealed that in order for an enterprise to get the fullest advantage of eTransformation, certain issues need to be addressed at the industry, and national level.

For the Ceramic Company studied, dealing with proper change management procedures was found to be an important aspect for the successful implementation of an eBusiness solution.

9. Conclusion

As the UNDP Human Development Report has described, people all over the world have high hopes that new technologies will lead to healthier lives, greater social freedoms, increased knowledge and more productive livelihoods [11]. eBusiness is a powerful tool, especially for developing countries, which will take advantage of the human capital intensive services and offer major opportunities to be competitive in a market dominated by developed countries. With the availability of the public network, open source software, e-business models, business process outsourcing as strategies, effective eTransformation, management of the eTransforming project is an important factor.

This paper shows that eTransformation models designed for the developed countries can be used for developing countries, but with modifications to incorporate the issues and problems specific to enterprises in developing countries. These issues should be addressed at four levels, namely, organisational, industry, national and global.

The 7E model shows a stage by stage process for enterprises to transform themselves by looking at the environmental factors and their own capabilities. The model also incorporates business strategies and eBusiness models to add a new dimension to the business to get competitive advantage in the eMarket. The 7E model gives the enterprises in any country enough flexibility and support to progress on their own pace to achieve success through eTransformation.

References

- [1]. Ginige A., Murugesan S., Kazanis P., *A Roadmap for Successfully Transforming SMEs in to E-Businesses*, Cutter IT Journal, May 2001, Vol 14.
- [2] UNDP(1999) Development Update 29,p.1
- [3] ECOSOC 2000 :Report of IT experts panel
- [4] Asia-Pacific Economic orporation (APEC) (1999) "SME Electronic Commerce-Study-Final Report" September 24, www.apecsec.org.sg/download/tel/SME_EcCommerce_study.exe
- [5] Schmögnerová B. Executive Secretary of UNECE - United Nations Economic Commission for Europe".
- [7] Ceramics Sector in Sri Lanka <http://competitiveness.lk/ceramics.htm>
- [6] Arunatileka, S. & Ginige, A. (2003) *Seven Es in eTransformation*, IADIS International Conference - e-Society, Lisbon, Portugal
- [8] UNCTAD (2002) *e-Commerce and Development Report*, United Nations Conference on Trade and Development, Geneva
- [9] Rayport, J. F. & Jawrski, B.J, (2001) *eCommerce*, New York, : McGraw-Hill, , p. 116.
- [10] Weill, P. & Vitale, M. (2001) *Place to Space – Migrating to eBusiness Models*, Harvard Business School press, Boston, Massachusetts.
- [11] UNDP, Human Development Report (2001), *Making New Technologies Work for Human Development*, UNDP, Oxford, UK
- [12]Shin, N. (2001) *Journal of Electronic commerce Research*, Vol2,No.4, : 164 - 171
- [13] Arunatileka, S. & Ginige, A. (2003) *Application of the Seven Es in eTransformation in the Manufacturing Sector*, eCahllenges International Conference –Bologna, Italy.



Web Site Visualisation as a User Navigation Aid

Shantha Jayalal Pearl Brereton Chris Hawksley
Department of Computer Science, Keele University
Keele, Staffordshire, ST5 5BG, United Kingdom

{shantha, o.p.brereton, chris}@cs.keele.ac.uk

Abstract

As e-society develops, web sites are containing increasing numbers of documents, often with complex interconnections. Existing tools are proving inadequate to enable users to navigate these web sites effectively. There are site maps to overcome this disorientation but these have limitations. In this paper we explore the use of a dynamic site map as a user navigation aid. We use an exponentially smoothed probability transition matrix for link prediction based on Markov theory and semantic clustering using lexical chains to obtain content similar pages. We introduce a prototype visualisation tool for overcoming the disorientation problem based on these principles. Applicability of link prediction & semantic clustering to other applications remains to be evaluated.

Keywords: Information Visualisation, Web site navigation, Disorientation, Site maps or Overview diagrams, link prediction, semantic clustering, lexical chains

1 Introduction

Despite its relatively short history, the Internet has become a powerful resource. It is forecasted that there will be 1 billion users connected to the Internet in the year 2003 compared to the 10 million users connected in the year 1997 [7]. Therefore the Internet today is one of the main infrastructure technologies of most organisations. Web sites have become the integration hubs for a wide variety of activities such as electronic commerce, online libraries and government services, rather than just disseminating information.

The large number of web pages on many web sites has raised navigation problems. The problems of disorientation or “becoming lost” on a web site are all too familiar to many of us. Often the only solution is to go back perhaps quite a long way to a

known point (such as a site’s home page). With a browser this may result in loss of the trace of some useful parts of one’s session [11].

We believe that improving user orientation can improve the user’s ability to navigate effectively. This is supported by other researchers who have shown that disorientation and ease of use are distinct, but negatively related and that lower perceived disorientation would result in higher performance [1].

Though most web pages provide some level of orientation data such as a title, this is a textual, subjective description from the author’s (not the reader’s) point of view. Most web sites do not offer visual orientation capabilities that can accommodate the site user’s point of view [10]. Many researchers have stressed the importance of improving site user orientation and have suggested the use of information visualisation techniques, in particular “site maps” or “overview diagrams” to address this issue [5, 10, 12, 14, 16]. Some site maps that are available in some web sites are static and texts based, and are often out of date, so site users do not have any interest in using them. Though some dynamic site maps are available in a few web sites, they have a highly complex structure [16].

The aim of this research is to improve the navigation within web sites by considering the patterns of access followed by users, the semantic contents of web pages and the topology of the site. A dynamic interactive web site visualisation model is proposed as a user navigation aid. The proposed model will be complementary to existing static text based web site maps.

Keele University Computer Science Department web site logs are being used in this study. The Department web site has a high volume of user access and the authors of this paper can obtain the logs for research purposes.

The structure of this paper is as follows. Definitions of disorientation, reasons for disorientation and methods to overcome disorientation are described in Section 2 and

Section 3 explains web site maps. Visualisation models are discussed in Section 4 and the proposed site map to aid user navigation is explained in section 5. Sections 6 and 7 describe the proposed prototype development process and future work respectively.

2 Disorientation

In this context disorientation has been defined as the tendency to lose one's sense of location and direction in a non-linear document [1]. Web site users sometimes feel lost, confused and overwhelmed when attempting to find information [10, 16].

"Where am I?" This is a question asked by most of us when we get dumped into the middle of a large web site by a search engine, often followed by, "what else is available on this site?" and "how do I get there?" [5]. As a web site user goes through a site, he or she has little or no sense of orientation of where he or she is, where he or she was, or where he or she is going.

Disorientation or the "lost in hyperspace" situation can cause users to become frustrated, lose interest, and experience a measurable decline in efficiency [1].

According to a survey by Zona Research Inc. (1998), 28% of Internet users reported that it was either somewhat or extremely difficult to locate specific products and 62% gave up looking for products on line. Lost clients can result in reduced revenues for organizations and ensuring that clients do not become disoriented should represent an important usability issue for designers [1].

Wasting client time through increased navigational complexity in a site is also an issue. Nelson and Sano (1995) estimate that about half a million dollars worth of user productivity is lost every time they add one more design element to Sun's home page [2].

2.1 Reasons for disorientation

We are interested in "lost in hyperspace" situations rather than problems deriving from issues such as network limitations, page loading speeds, and network bandwidth.

Web sites pose many problems: they are often very large, complex, change rapidly, and are very diverse in both content and form of information, and more importantly, are often ill-structured [11]. Two reasons why web site users are disoriented when navigating web sites are [3, 5, 10, 12, 15, 16]:

- Most complex web sites do not provide a clear conception of relationships within pages of the web site to the user. This is because by nature, and by poor design, web sites lack an organizational paradigm. A printed document has a clear order of presentation. On a web site anything can be connected to anything and intended ordering is more difficult to convey. As a result, the user may not know his or her present location in the web site, and finds it difficult to decide where to look next within the site.
- Web browser interfaces do not provide a concise summary of the outgoing links from a page. Also there is no way of finding incoming links, since there is no central link database to consult.

2.2 Overcoming disorientation

Several methods have been suggested to alleviate navigational problems. Some of these are [3]:

- New Internet browsing applications could be developed to supersede current browsers. This approach has several drawbacks including the complexity of browser design, and the need to deal with the large amount of media types available on Internet.
- The design of a web site could be improved to support more effective navigation. Such improvements can be made in the areas of graphic design, content presentation, and the integration of user-centred navigation aids, such as overview maps.

In this paper we are interested in developing techniques to support user navigation rather than addressing the issues of new browser applications.

Well-designed navigation systems can help to solve the disorientation problem experienced in web sites [1], but the problem of global navigation across the Internet may never be solved [10]. It is therefore left to the designers of individual web sites to provide navigation aids with clear information about the nature of the information space being traversed and available paths through it [11, 16].

We believe graphical representations have much to offer in this context. Diagrams could be used as an interface for navigating the web [12]. One of the ways in which web site designers are trying to address this problem is by providing what is commonly called "site maps" or "overview diagrams"[5, 10, 12, 14, 16].

The idea of a web site map is based on the geographical metaphor of the map. Overview diagrams or site maps can be used to visualise the structure and contents of the underlying information space so that users see where they are, what other information is available and how to access it [15]. The development of these maps is commonly known as “web site mapping”. However, the art and science of creating useful web site maps is still in its infancy [5, 10, 14].

We propose to reduce disorientation within web sites by presenting a visualisation of the path followed by a user, the probable next move from the current page and semantically similar pages to the current page.

3 Site maps or Overview diagrams

There are a number of different types of web site maps ranging from simple text indexes and “table of contents” to sophisticated graphical representations. However, good site maps are those that satisfy at least the following two requirements [5]:

- The viewer must be able to comprehend what the diagram represents
- The viewer must be able to perceive his or her own location within the diagram

Most site maps are constructed manually and are static, involving studying the contents of a web site manually or semi-automatically and then determining the grouping of web pages and giving them some pictorial or visual representation. Dynamic maps are produced “on the fly” when the web sites visited. In dynamic mapping the system automatically analyses the structure of the web site. This makes it possible to always get the most recent web site structure for visualisation. A dynamic web site map may appear differently from time to time depending on the current contents of the web site, user’s current location (web page) and the path followed by the user of the web site [10].

3.1 Pros and cons of site maps

While the main purpose of site maps is to improve user orientation, there are some other benefits. Site maps support the user’s mental model of how the site space is structured, showing the breadth and depth of the site and acting as a visual surrogate for short-term memory. This mental model can be a powerful aid to reduce learning time and improve overall performance [16].

Site maps reduce the cognitive overhead, the additional effort and concentration necessary to maintain several tasks. In the absence of visual aids, users will have to make extra effort to create some kind of mental picture/map about the sources of information and their relations among the sources [4,5,10,17].

Predictably, there are problems with using site maps, which do not provide inherent clues to the navigation of the site. There is also a risk of using site maps as a “bubble gum” approach to fixing poor site design. Speed, complexity and maintenance problems have been attributed to poorly designed site maps [16].

3.2 Characteristics of site maps

Durand and Khan [5] have identified a set of requirements for web site maps as the following, reflecting on problems faced by web site users, authors and administrators.

- i show a high level view of the web site structure
- ii show where the web site user is: his/her current page
- iii show which pages the web site user has visited and pages that remain unexplored
- iv show where the web site user can go from his/her current page
- v distinguish peer group relationships: show the pages that are in the same group as the current page
- vi show the path followed (visited pages in order) to arrive at the current page
- vii show how many other web site users have viewed the current page
- viii distinguish interesting pages: show the pages that would be of interest to the user depending on the contents of the page and the user’s interest
- ix display all this information in a very small space: screen size is limited to 600 x 400 pixels
- x display the map in a web browser without requiring additional software or lengthy download time

Durand and Khan [5] have focused on features (i) to (v) and features (ix) to (x). Cockburn and Jones [3] have proposed navigation support functions for a graphical site to assist in user navigation through complex information space. They include:

- switching facility (between the browser and visually represented site map)
- link preview (showing all the available links out of the page)
- visualisation filtering (allowing the user to control the amount of detail, type, and style of the map)

4. Visualisation models for web site navigation

There are a number of web site visualisation models available at present for various purposes. Models such as VisVip by Cugini and Scholtz[8], EbizLive by Eick [6], Starfield by Hochheiser and Shneiderman [9], WebViz by Pitkow and Bharat [18] are intended mainly to visualise web site usage patterns and are very useful for site authors, site operators, marketing and business managers. However, they do not provide information from a site user's perspective.

The visualisation model reported by Wong et. al. [20] shows the data generated by a particular page in relation to only those pages that are relevant to it. This model visualises the statistics produced by server logs while incorporating the site structure, and allows a user to narrow their attention to a particular sub-branch of interest. The metaphor chosen for this site layout view is that of a Windows type directory tree listing. Selecting a page in the site layout view activates the zoomed view. That page will then appear in the zoomed view, along with accompanying pages below it in the tree. Individual statistics on each page will then be displayed in the manner of bars appearing on the corresponding page.

This site layout view is like a hierarchical tree, even though web sites are closer to networks in terms of their structure. Thus users might be uncomfortable with this strange new metaphor. In addition, the site layout view and zoomed views are not capable of displaying user-accessed pages in order.

MAPA is a dynamic Java-based hierarchical site map application [5]. It is a powerful way to visualise a web site's organisation, content and scale. MAPA displays coloured boxes, where each box represents a page on the web site. A white box with a red base shows the user's current location. Green and blue boxes represent the child pages and grandchild pages of the current page respectively. When the user moves the cursor over a box, the title and URL are displayed. If the user clicks on a box, the diagram rearranges itself with the clicked

page as the current location. If the user double-clicks, that page is opened in the web browser.

Site users must carefully position the cursor over box after box, reading the titles and URLs of each page to find a particular page with MAPA. This is a major usability problem when there are dozens of green and blue boxes. Also most web sites are not purely hierarchical, as is assumed for site maps in MAPA. Another problem of MAPA is the non-availability of dynamic site usage information.

The above mentioned visualisation models are limited solutions to web site user disorientation problems. In particular they do not perform well on large sites. Their limitations relate, variously, to poor usability, non-availability of paths followed, non-availability of relationships among pages and inability to predict next possible moves.

Many current web sites have thousands of pages and links between them. Because of this complex structure there is a need for new visual models to improve web site user orientation.

5 Proposed site map as a user navigation aid

Our hypothesis is that "web site user disorientation can be reduced by dynamic navigation aids which include visualisation of the user current page, semantically similar pages to the current page, link prediction from the current page and the path followed by the user so far". We propose the use of a dynamic site map or overview diagram as a user navigation aid, which will be able to display information relating to the following:

- (i). a high level view of the web site structure with
 - current user location (where am I?)
 - user history (where have I been so far?) in terms of visited pages and the path followed
- (ii). a localised view related to the current page
 - the most probable next move considering the path followed so far (link prediction)
 - clusters of pages which are similar to the current page
 - other possible destinations from the current page

The proposed site map should have the following features:

- possibility of exploring any page if it is seen on the map

- browser independence
- switching facility (between the browser and visually represented site map)
- display of all this information in a relatively small limited space (trade off between the size and content)
- possibility to display the map in a web browser without requiring additional software or lengthy download time

The user can be presented with an option to generate a site map from anywhere in the web site. The focal node of the site map represents the user's current page. A page will be opened through a web browser by double clicking on any node within the diagram.

Our link prediction is based on Markov theory [19] and semantic clustering using lexical chains to assess the content of similar pages is employed. These are explained in the following sections.

5.1 Link prediction

Probability transition matrices based on Markov theory have long been used for prediction of future activities in the field of Operations Research. In this paper a probability transition matrix (Q), is determined for link prediction.

A sample link graph illustrating the probabilities of selecting web pages in the Keele Computer Science web site is shown in Figure 1. For example P_{12} is the probability of selecting the link to page 2 from page 1. These probabilities can be calculated by mining the web logs of the particular web site.

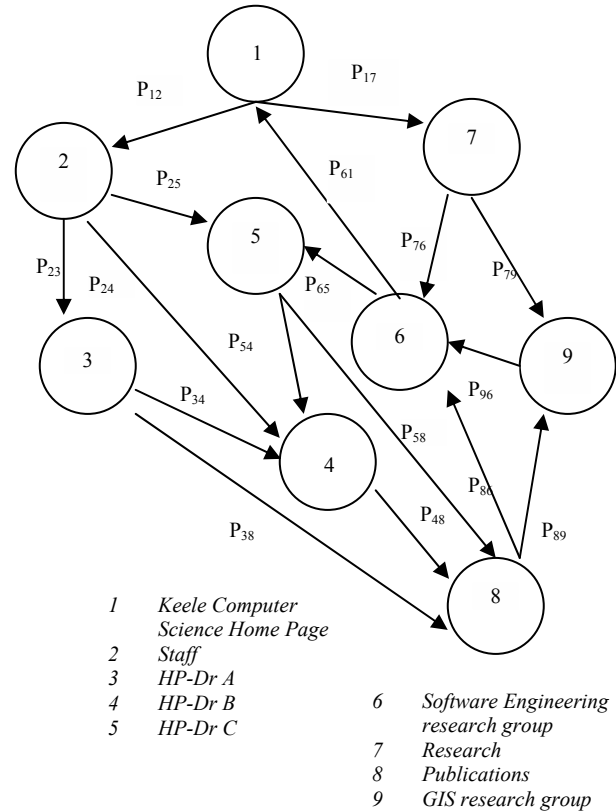


Figure. 1: Link graph of a sample web site

These probabilities can be represented as a transition probability matrix (Q) as shown in Figure 2.

Page No	1	2	3	4	5	6	7	8	9
1		P_{12}					P_{17}		
2			P_{23}	P_{24}	P_{25}				
3				P_{34}				P_{38}	
4								P_{48}	
5				P_{54}				P_{58}	
6	P_{61}				P_{65}				
7						P_{76}			P_{79}
8						P_{86}			P_{89}
9						P_{96}			

Figure 2: Transition probability matrix (Q) for the link graph in Figure 1

And, for any given page i which contains at least one out going link:

$$\sum_{j=1}^9 P_{ij} = 1$$

Where $P_{ij}=0$ if there is no link from page i to j , otherwise $0 \leq P_{ij} \leq 1$, for $i=1..9$

It can be argued that the most recent data possible should be used in Q in order to provide the best predictions. However a more reliable overall set of values for Q requires data collected over a longer period in order to reflect the long-term usage of the site as whole. So we propose to form a Q using the single exponential smoothing average method. Consider several transition matrixes, Q_t , Q_{t-1} , $Q_{t-2} \dots Q_{t-n}$, that are formed in different time periods. The most recent transition matrix at time t , Q_t , is made most influential in calculating the single exponential smoothed moving average Q over these n periods given by:

$$Q = \mu Q_t + \mu(1-\mu)Q_{t-1} + \mu(1-\mu)^2 Q_{t-2} + \mu(1-\mu)^3 Q_{t-3} + \dots + \mu(1-\mu)^n Q_{t-n}$$

where μ is a constant and $0 \leq \mu \leq 1$

Here greater weight is given to more recent transition probabilities and n previous transition probabilities are taken into account.

Based on Markov theory, Sarukkai [19] proposed to use the link history vectors ($L_0, L_{-1} \dots L_{-m+1}$) of a user with the transition probability matrix to calculate vector N , for the probability of each page to be visited in the next step as follows:
 $N = a_1 \times L_0 \times Q + a_2 \times L_{-1} \times Q^2 + \dots + a_m \times L_{-m+1} \times Q^m$

Where $a_1, a_2 \dots a_m$ are the weights assigned to the history vectors. Normally, $1 > a_1 > a_2 > \dots > a_m > 0$, so that the closer the history vector is to the present, the more influence it has to the future. His/her visiting history sequence is represented as m number of pages from 0 to $m-1$. Each and every followed link of the link history is represented as a vector with a probability 1 at that state for that time (denoted by $L_0, L_{-1} \dots L_{-m+1}$). Vector, L_0 represents the current page of the user.

5.2 Clustering

As proposed at the beginning of Section 5 we use clustering to show peer group relationships in the site map. The similarity between web pages is calculated based on their semantic content. The

semantic content of web pages can be represented using lexical chains.

A lexical chain is a sequence of semantically related words in a text [13]. In their work, Morris and Hirst have demonstrated that the structure of the lexical chains in a document corresponds to the structure of the document itself. Green [8] has successfully used lexical chains to build hypertext links within a newspaper article considering its paragraph semantic similarity. In his lexical chaining process he has not considered the strength of a lexical chain.

The lexical chains in a text can be identified using any lexical resource that relates words by their meaning. We use the WordNet lexical database [21] for our lexical chaining process. WordNet is an electronic lexical knowledge base developed at Princeton University. The WordNet database is composed of synonym sets or synsets. Each synset contains one or more words that have the same meaning. A word may appear in many synsets, depending on the number of senses that it has. Synsets can be connected to each other by several different types of links that indicate different relations. For example, two sets can be connected by a hypernym link, which indicates that the words in the source synset are instances of the words in the target synset.

The WordNet database is divided into four data files containing data for adjectives, adverbs, nouns, and verbs respectively. An index file is associated with each data file.

Each index file (see Figure 3) is an alphabetised list of all the words found in WordNet in the corresponding part of speech. Associated with each word is a list of byte offsets (synset offsets) in the corresponding data file identifying each synset containing the word. A data file (see Figure 4) contains a set of tuples. Each tuple corresponds to a synset, which contains words of similar meaning that refer to a common semantic concept. As words can have more than one meaning, a word may be present in more than one synset.

Index File Format
lemma pos synset_cnt p_cnt [ptr_symbol...]
sense_cnt tagsense_cnt synset_offset
[synset_offset]

taxonomy n 3 2 @ ~ 3 0 06916721
05212711 00770475
taxophytina n 1 2 @ #m 1 0 09609154
taxopsida n 1 2 @ #m 1 0 09609154
taxpayer n 1 2 @ ~ 1 1 08756381

Figure 3: Index File Sample

Data File Format									
<i>Synset_offset</i>	<i>lex_filename</i>	<i>ss_type</i>	<i>w_cnt</i>	<i>word</i>					
<i>lex_id</i>	[<i>word lex_id...</i>]	<i>p_cnt</i>	[<i>ptr...</i>]	[<i>frames</i>]					
<i>gloss</i>									
08756381	18	n	01	taxpayer	0	002	@		
08531123	n	0000	~	08607752	n	0000			
someone who pays taxes									

Figure 4: Sample tuple from data file

In order to form lexical chains, semantic contents from all the web pages of our target web site are extracted. Then the stop words (vague high-frequency words e.g., one, two, dozen, right, ok, is, am, are, were etc) are removed. The final word list will contain all the unique words with their frequencies from each and every web page of the site. Calculating the relationships between each and every word forms lexical chains. We have defined three kinds of relationships between words as extra-strong (Figure 5), strong (Figure 6) and medium strong (Figure 7). Here double ovals correspond to words while single ovals correspond to synsets. The synset common to two words is highlighted.

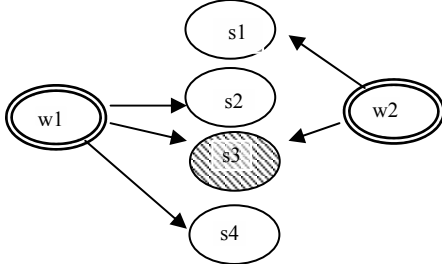


Figure 5: Extra-strong relationship between word_1 and word_2

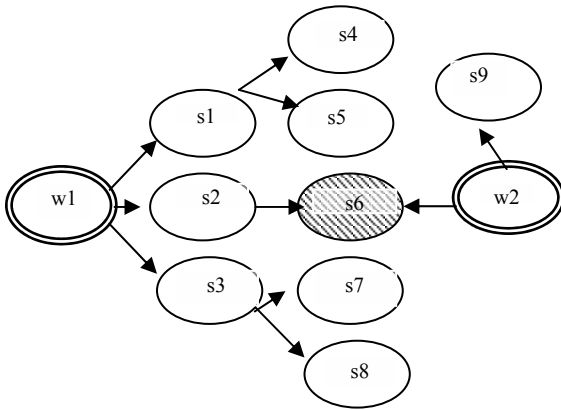


Figure 6: Strong relationship between word_1 and word_2

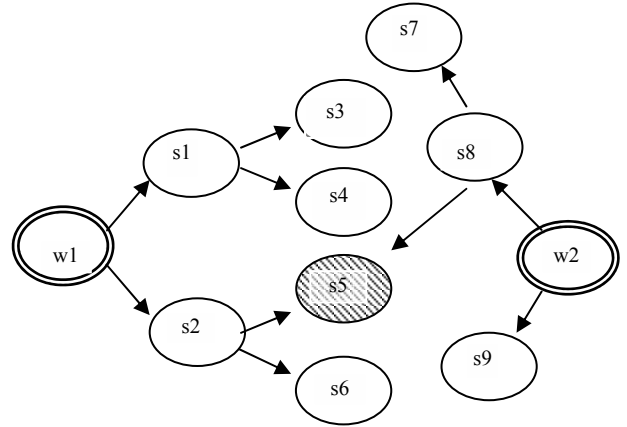


Figure 7: Medium strong relationship between word_1 and word_2

Extra strong relations have the highest weight of all relations. A strong relation has a lower weight than any extra-strong relation and a higher weight than any medium-strength relation. A word will be included in the most suitable lexical chain by calculating the semantic relationship between the word and each and every other word in the lexical chain.

The strength of the relationship between words represents the Lexical Semantic Distance (LSD) between words. Consider two words word_1 and word_2. The relationship strength is:

Extra Strong => semantic distance = 1
 Strong => semantic distance = 2
 Medium Strong => semantic distance = 3

When calculating the lexical chains, the strength of the relationship between words is used in order to calculate the chain strength.

5.2.1 Chain Strength Analysis

Some lexical chains are stronger than others. Strong chains are more likely to have semantic correspondence to the text than a weak one [13]. Morris and Hirst have identified the strength of a lexical chain i as a function of:

- Reiteration (R_i) – the more repetitions, the stronger the chain
- Density (D_i) – the denser the chain, the stronger it is
- Length (L_i) – the longer the chain, the stronger it is

R_i , D_i and L_i can be defined as:

$$R_i = \frac{TNW_i - TNUW_i}{TNW_i}$$

$$D_i = \frac{TNW_i}{TNW_ALL} \quad L_i = \frac{TNUW_i}{TNUW_ALL}$$

where

TNW_i - Total no of words in the chain i

TNUW_i – Total no of unique words in chain i

TNW_ALL - Total no of words in all the chains

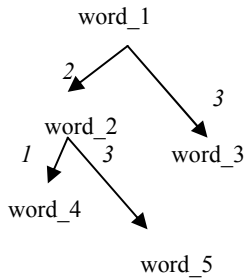
TNUW_ALL - Total no of unique words in all the chains

The Lexical Semantic Distance (LSD) of a lexical chain is also calculated in order to obtain the strength of a lexical chain. The LSD of the lexical chain i (LSD _{i}) is the total of the semantic distances between words of the lexical chain i . The Average LSD of the lexical chain i (ALSD _{i}) is obtained by dividing the LSD by total number words of the lexical chain i . The lower the ALSD, the stronger the chain.

Consider two lexical chains, LC1 and LC2, and their lexical semantic distances (represented as lexical semantic trees) as given in Figure 8 and Figure 9 respectively. A lexical semantic distance tree represents the strength and the relationship between each and every word of a lexical chain. The numbers (1, 2 or 3) on the arcs between words represents the strength between words (whether extra strong (1), strong (2) or medium strong (3) as defined above).

LC1 (word_1, word_2, word_3, word_4, word_5)

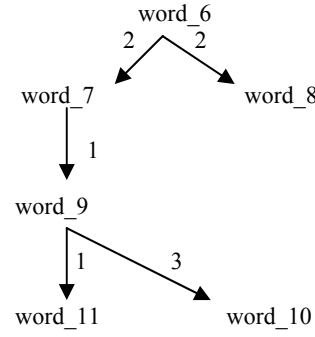
LC2 (word_6, word_7, word_8, word_9, word_10, word_11)



$$LSD(LC1) = 2 + 3 + 1 + 3 = 9$$

$$ALSD(LC1) = 9/5 = 1.8$$

Figure 8: Lexical semantic distance tree from LC1



$$LSD(LC2) = 2 + 2 + 1 + 1 + 3 = 9$$

$$ALSD(LC2) = 9/6 = 1.5$$

Figure 9: Lexical semantic distance tree from LC2

We define overall strength of a lexical chain i (StLC _{i}) as follow:

$$StLC_i = \frac{\alpha R_i + \beta D_i + \delta L_i}{ALSD_i}$$

where

$$0 \leq (\alpha, \beta, \delta) \leq 1$$

Document Density (DD) of Lexical Chain c (LC _{c}) in document i (DD _{ci}) is defined as:

$$DD_{ci} = \frac{TNW_{ci}}{TNW_i}$$

Where

TNW _{ci} - Total no of words from chain c that appear in document i

TNW _{i} - Total no of words in document i

Document density between lexical chains and documents for our sample web site are represented as a matrix (a document density matrix) as in Figure 10.

The similarity between two chain density vectors can be calculated using the Dice association coefficient ($S_{Dice}(p, q)$) by incorporating the strength of the lexical chain. Here greater numbers indicate a greater similarity.

For any two documents p and q , incorporating chain strength, $S_{Dice}(p, q)$ is given by :

$$S_{Dice}(p,q) = \frac{2 \sum_{i=1}^n StLC_i(DD_{ip} \cdot DD_{iq})}{\sum_{i=1}^n (StLC_i \cdot DD_{ip})^2 + \sum_{i=1}^n (StLC_i \cdot DD_{iq})^2}$$

where

n is the no. of lexical chains

The similarity matrix can be obtained after calculating $S_{Dice}(p,q)$ for all the documents of our sample web site as in Figure 11. Higher similarity coefficients represent higher similarities between pages based on their semantic content.

6. Prototype

An illustration of our prototype site map (a Java applet) is shown in Figure 12 for the web site in Figure 1. A particular user has started a journey from Keele Computer Science Home Page and is now at the home page of Dr C (HP-Dr C).

The path followed by the user starting from Keele Computer Science Home Page, is:

Keele Computer Science Home Page → Research → GIS research group → Software Engineering research group → Keele Computer Science Home Page → Staff → HP-Dr C

Therefore, now, his or her current page is HP-Dr C. A sample site map of our prototype focussing on the home page of Dr C is shown in Figure 12 where circles represent web pages and lines represent links between pages. The size of the circle represents the popularity of the page based on the number of hits recorded in the log file. The current page (HP-Dr C), marked as page 5, is close to the centre of the site map.

The thickest arrow from node 5 represents the most probable next move from the current page depending on the user link history and the single exponentially smoothed transition matrix (from HP-Dr C to the Publications page).

LC _i	StLC _i	Document (d)								
		1	2	3	4	5	6	7	8	9
1	StLC ₁	DD ₁₁	DD ₁₂	DD ₁₃	DD ₁₄	DD ₁₅	DD ₁₆	DD ₁₇	DD ₁₈	DD ₁₉
2	StLC ₂	DD ₂₁	DD ₂₂	DD ₂₃	DD ₂₄	DD ₂₅	DD ₂₆	DD ₂₇	DD ₂₈	DD ₂₉
3	StLC ₃	DD ₃₁	DD ₃₂	DD ₃₃	DD ₃₄	DD ₃₅	DD ₃₆	DD ₃₇	DD ₃₈	DD ₃₉
4	StLC ₄	DD ₄₁	DD ₄₂	DD ₄₃	DD ₄₄	DD ₄₅	DD ₄₆	DD ₄₇	DD ₄₈	DD ₄₉

Figure 10: Document Density Matrix

Document	1	2	3	4	5	6	7	8	9
1		$S_{Dice}(1,2)$	$S_{Dice}(1,3)$	$S_{Dice}(1,4)$	$S_{Dice}(1,5)$	$S_{Dice}(1,6)$	$S_{Dice}(1,7)$	$S_{Dice}(1,8)$	$S_{Dice}(1,9)$
2			$S_{Dice}(2,3)$	$S_{Dice}(2,4)$	$S_{Dice}(2,5)$	$S_{Dice}(2,6)$	$S_{Dice}(2,7)$	$S_{Dice}(2,8)$	$S_{Dice}(2,9)$
3				$S_{Dice}(3,4)$	$S_{Dice}(3,5)$	$S_{Dice}(3,6)$	$S_{Dice}(3,7)$	$S_{Dice}(3,8)$	$S_{Dice}(3,9)$
4					$S_{Dice}(4,5)$	$S_{Dice}(4,6)$	$S_{Dice}(4,7)$	$S_{Dice}(4,8)$	$S_{Dice}(4,9)$
5						$S_{Dice}(5,6)$	$S_{Dice}(5,7)$	$S_{Dice}(5,8)$	$S_{Dice}(5,9)$
6							$S_{Dice}(6,7)$	$S_{Dice}(6,8)$	$S_{Dice}(6,9)$
7								$S_{Dice}(7,8)$	$S_{Dice}(7,9)$
8									$S_{Dice}(8,9)$
9									

Figure 11: Document Similarity Matrix

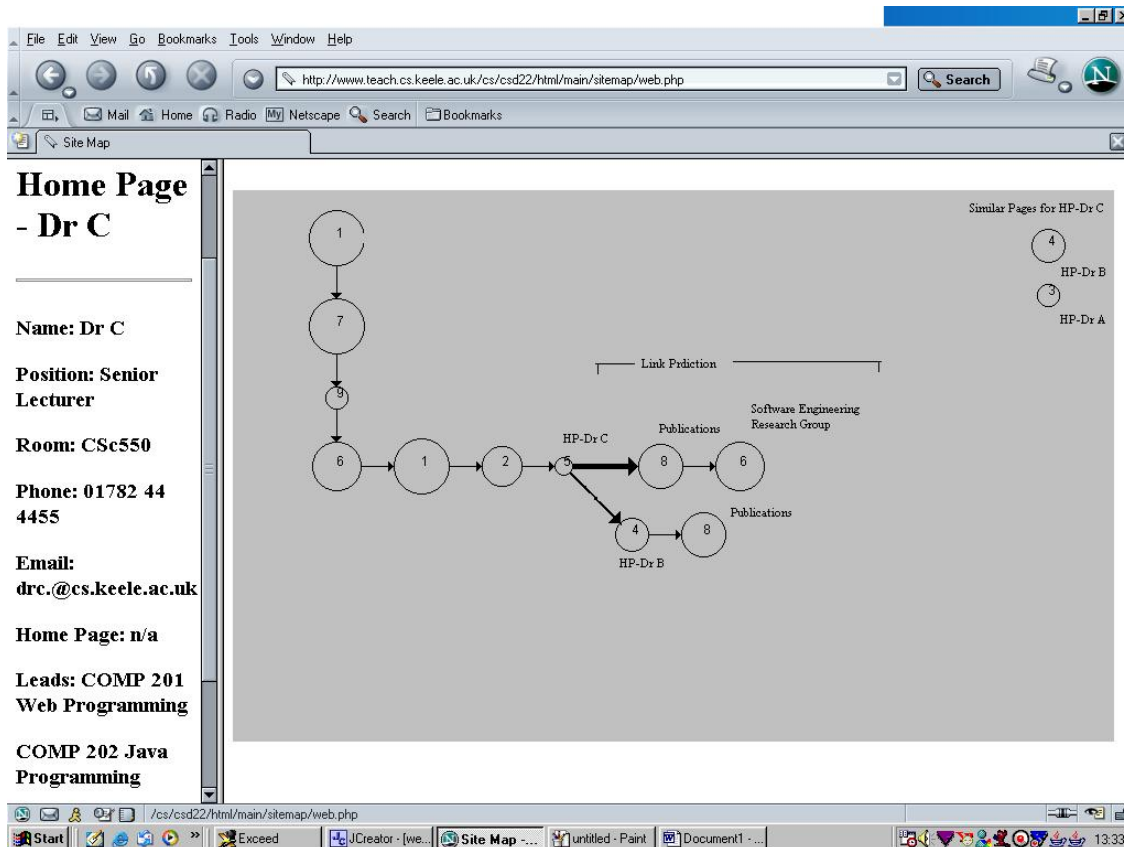


Figure 12: Prototype site map

The other thick arrow represents the 2nd most probable next move.

The other pages in the same cluster as the current page (HP-Dr B, HP-Dr A) can be seen in the top right hand corner of the site map and represent the peer group relationship with the current page (HP-Dr C).

The semantic similarities between web pages of the Computer Science department are being calculated using the above-mentioned semantic clustering method. We have developed in Java, a web crawler and a lexical chain generator incorporating the WordNet lexical database, which is now being tested.

The access logs generated by the Keele University web server will be pre-processed in order to form transition probability matrices for different time periods, which will then be combined into one matrix using the single exponential smoothing average method. The most likely next move of the user will be obtained by using the Sarukkai method [19] as outlined above.

7. Conclusion & future works

Present day web sites have thousands of pages and links between them and this leads to high levels of user disorientation. Because of this complex structure of web sites there is a need for new visual models to improve web site user orientation. This paper presents a method for overcoming disorientation using dynamic site maps.

We propose to carry out a number of empirical case studies using our prototype in order to deduce an appropriate value for the transition matrix constant and hence determine the accuracy of link prediction. The usability of the site map needs to be evaluated by a larger user group. We plan to select a group of subjects including students and staff in our university, and people from outside the university to evaluate the usability of our site map.

The applicability of the link prediction methodology and semantic clustering to other applications such as web server HTTP request prediction, adaptive web navigation, and automatic tour generation remains to be evaluated.

References:

1. Ahuja S. J., Webster J., (2001). "Perceived disorientation: an examination of a new measures to assess web design effectiveness", *Interacting with Computers*, 14, pp 5-29
2. Cockburn A., Jones S., (2000). "Which way now? Analyzing and easing Inadequacies in WWW navigations", *International Journal Human-Computer Studies*, 45, pp 105-129, Academic Press, December
3. Cockburn A., Jones S., (1997). "Design Issues for World Wide Web Navigation Visualisation Tools", In proceedings of the RIAO'97: The 5th Conference on Computer-Assisted Research of Information, McGill University, Montreal, Quebec, Canada, 25th-27th June, pp 55-74
4. Cugini J., Scholtz J. (1999) "VISVP: 3D Visualisation of Paths through Web Sites", *Proceedings of the International Workshop on Web Based Information Visualisation (WebVis'99)*, pp 259-263
5. Durand D., Kahn P., (1998) "MAPATM: a system for inducing and visualizing hierarchy in websites", 9th ACM Conference on Hypertext & Hypermedia, June
6. Eick G. S. (2001) "Visualising Online Activity", *Communications of the ACM*, August, Vol. 44, No. 8
7. Fensel D., Musen M. A., (2001) "The Semantic Web: A Brain for Humankind", *IEEE intelligent Systems*, March/April
8. Green, S. J. (1999). Building hypertext links by computing semantic similarity, *IEEE Transactions on Knowledge and Data Engineering*, vol., 11, No 5, September/October
9. Hochheiser H., Shneiderman B. (1999) "Using Interactive Visualisations of WWW Log Data to Characterise Access Patterns and Inform Site Design", A revised version appears in *ASIS'99 Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, Annual Conference October 31-November 4, Vol. 36, pp 331-344
10. Hung L. M., "Information Visualization in Web Site Mapping: a Survey", www.iiis.org/sci/program2000.htm
11. Inder R., Kilgour J., Lee J., (1998) "Automatic generation of Diagrammatic Web Site Maps", *ACM*
12. Lai W., Danaher M., (1999) "An Approach to Graph Layout to Assist in Web Navigation", 3rd International Conference on Computational Intelligence & Multimedia Applications", 23-26, September, New Delhi, India
13. Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17, No 1, pp. 21-48
14. Mukherjee S., Foley J. D., (1995) "Visualizing the World-Wide Web with the Navigational View Builder", *Computer Networks and ISDN Systems*, 27(6), pp 1075-1087
15. Mukherjee S., (2000) "Information Visualization for Hypermedia Systems", *ACM*
16. Pilgrim C. J., Leung Y. K. (1999) "Designing WWW Site Map Systems", *Proceedings of the International Workshop on Web Based Information Visualisation (WebVis'99)* in conjunction with DEXA'99 Tenth International Workshop on Database & Expert Systems Applications, pp 259-263
17. Pilgrim C., Leung Y., (1996) "Applying Bifocal Displays to Enhance WWW Navigation", 2nd Australian WWW Conference, Southern Cross University, 7- July
18. Pitkow J. E., Bharat K. A., (1994) "WebViz: A tool for world wide web access log analysis", 1st International World Wide Web Conference, Geneva
19. Sarukkai R. R. (2000) "Link prediction and path analysis using Markov chains", *Computer Networks* 33, pp 377-386
20. Wong B., Marsden G. "Using Access Information in the Dynamic Visualisation of Web Sites", www.cs.uct.ac.za/Research/CVC/Techrep/CS00-18-00.pdf
21. WordNet 1.7.1 Reference Manual, <http://www.cogsci.princeton.edu/~wn/doc.shtml>



Data Protection Law an E-Business and E-Government Perception

Prathiba Mahanamahewa
Ph.D Research Scholar in IT Law

Abstract

This paper examines the extent to which the basic principles of data protection laws may be read into provisions in human rights treaties proclaiming a right to privacy. The problems caused by the difference between the USA and EU approaches to data protection are highlighted. The paper documents some of the legislative developments mean for E-Business and the link between data captured, stored and processed into information and the resulting effect on privacy is important. This article further argues in Sri Lanka the Information Privacy should be constitutionally protected and the law should be based on a EU model considering EU derivatives (1995/EC/46).

1. Introduction

According to the United Nations Conference on Trade and Development (UNCTAD), E-commerce and Development report of 2002, revealed that the global number of Internet users is expected to reach 655 million by the end of 2002. (Compared to 500 million at the end of year 2001.) An emerging feature of this report is that during last year a growing share of new Internet users was found in developing countries, accounting for nearly a third of new Internet users worldwide. In Sri Lanka during 2002 there was nearly a 25 percent increase of the Internet usage, compared to the year 2001. Internet usage is having an annual rise of about 30 percent, which is equivalent to about 2.5 percent of the global population. The report further states that the world's largest e-commerce market has experienced a mild recession in 2002 [1]. When analyzing reasons for this decline, several surveys conducted in New Zealand, Australia and U.S.A, revealed that there is a great fear among the consumers about feeding their personal information to computers for e-business. Is their personal data adequately protected and well secured? If this fear continues in future it may badly affect the e-business development. This paper discusses the law relevant to data protection giving special attention to Sri Lanka.

2. Why do Countries Enact Data Protection Legislation?

The most commonly stated motive for data protection legislation is to protect individual privacy from being compromised by computerization and to provide a framework for finding a balance between the interests of the individual, the data user, and the community at large. The 1981 Council of Europe Convention on Data Protection forms the basis of all national legislation within Europe. The U.K. Data Protection Act 1998 was with a view of protecting private individuals from the threat of the use of erroneous information regarding them, or the misuse of correct information about them held in computers. This satisfied the Council of Europe Convention on Data Protection, and enabled the U.K. data processing industry to participate freely in the European market. India is ready with their draft of the data protection act and is planning to enact legislation in the coming winter session of the parliament.

It has been revealed that the countries without a data protection law may lose business especially from Europe because the EU Directive on the Protection of Personal Data prohibits the transfer of personal data to non-EU countries that do not meet the European "adequacy" standard for data protection. As a result, this Directive also places important burdens on Sri Lanka and other countries that collect personal data online. The absence of data privacy legislation in India has proved to be a handicap to Europe and U.S.A to carry on business processes with India and Indian companies.

3. What is Data Protection?

Providing a definition for data protection is as difficult as finding the reason for its existence. It is the legal protection of individuals with regard to automatic processing of personal information relating to them. Therefore not only should there be a clear balance between data users and data subjects, but also a scale of values attached to any individual [2].

4. Why Data Protection Law Needs at this Hour?

The use of Internet and e-mail by the Sri Lankan government sector and the private sector has been rapidly increased over the past few years. They have launched an e-government project to deliver accurate and prompt services to the public. Most of the government institutions in Sri Lanka are connected to e-mail and Internet associated with a high computer use. Sri Lanka has to face these new challenges posed by the electronic revolution to protect individual privacy of processing their personal information. In many countries including Sri Lanka, laws have not kept up with the technology, leaving significant gaps in protection. In other countries, law enforcement and intelligence agencies have been given significant exemptions. Therefore the mere presences of law may not in fact provide adequate protection.

5. What is E-Government?

E-government is all about government agencies working together to use technology so that they can provide individuals and the business sector with better government services and information. It is not a massive Information Technology (IT) project. Much of it is about establishing more effectively common standards across government, delivering services, and providing better ways for agencies to work together, using technology. E-government can enhance the citizen's access to government information and services, and can provide new ways to increase citizen participation in the democratic process and in building a knowledge-based economy with sustained prosperity.

E-government makes it easier to do business with government. It also makes it cheaper. In Australia it has been estimated that it can cost as little as \$1-7 each time you use a service online. This compares with \$2-200 to deliver the same service over the counter, by mail or telephone, or even by sending out a brochure.

Businesses will find it much easier to work with government organizations than they do today. They will specially notice the reduced cost of dealing with government. Business people will also be able to find the right information and regulations. They will also be able to conduct the related transactions quickly and in an integrated way. By encouraging use of the Internet, e-government will indirectly foster opportunities for business to develop their online services and would be a real boost for a business with a small workforce.

6. What will E-Government Mean for Small and Large Business Ventures?

A significant benefit for business will arise from the participation of government in the information economy. While the focus of this program is on taking government online, the effect will be to stimulate and to move a critical mass of Sri Lanka business services online. This would make a positive impact on for our ability to operate in the global economy.

7. What is E-Business?

The convenience, availability and worldwide reach of E-business enhances existing business or creating a new virtual business. IBM defines e-business as "a secure, flexible and integrated approach to delivering differentiated business value by combining the systems and processes that run core business operations with the simplicity and reach made possible by Internet technology". E-commerce is just one aspect of e-business like e-franchising, e-mailing, e-marketing, e-business, the Internet, and globalization all depending on each other. The more global players exist the more e-business would want to play its part. The more e-business is on-line, more people will be attracted to get direct Internet access. The more people are online more global players will arise.

8. The Relationship Between E-Government, E-Business and Data Privacy

Online business or online delivering public services, means feeding personal data to the computer. These data are transmitted from one place to another place. When this is being done the data collectors, service providers or data registers are involved in this activity. The personal data of customers and citizens should thus be protected by law without keeping any room for manipulation, possible misuse or unauthorized disclosure to a third party.

9. Threats to Personal Privacy

Consumer awareness about privacy is increasing, particularly among Internet users. Sooner or later, consumers will demand that their privacy be respected. This may require some modification to business practices and customer service and may even involve costs not previously incurred. Even American big business has accepted that privacy is a concern, which must be addressed. All the public surveys conducted by, and for big business in America, showed a lack of confidence in that consumer's personal information may not be protected if they entered into transactions on the Internet.

Privacy concerns have thus been clearly identified as a barrier to the development of e-business [3].

Firstly, personal information that an individual would prefer not to disclose to others, can be obtained from imprints left by identifiers on the hard drive of a computer. For instance, in registering Microsoft Word, an identifier was placed on the hard drive that could have permitted Microsoft to track all movements on the Web. Although Microsoft changed the registration system, an identifier is now made through registration of Microsoft Media Player, as well as through other software systems.

Secondly, web bugs can similarly disclose personal information that many of us would prefer to keep confidential. Web bugs are images embedded in a web page that can transmit information to a remote computer when the page is viewed. The remote computer can track which computer accesses which page. The Web bug is quite a recent innovation. They are also known as clear GIFs, or 1 x 1 GIFs. A web bug is a tiny graphic, included in a web page or e-mail message, used to identify who or how many people are viewing the material. They can be placed in the image tags of the underlying HTML code of the page and they can also be placed in HTML enabled e-mail messages. For instance, Toys-R-Us.com used a tracking device to compile information about online shoppers. After August 2000 when this practice was discovered, the site discontinued the practice. This apparently violated Toys-R-Us' privacy policy.

Thirdly, an Internet Service Provider (ISP) is a gateway to the Internet. ISPs hook up a personal computer or system of computers to the Internet. ISPs can divulge a host of information about an individual, including name, address, and credit card. They can recapture e-mail that was sent through their services. ISPs can also recapture session information, such as the URLs visited by a user, through its service. At times they have disclosed private information about individuals, leading to embarrassment and adverse employment consequences

Fourthly, cookies are small text files placed on an Internet user's computer when a website is accessed. They contain information sent by the server to the user's browser. If desired, a web user can sometimes view cookies in the source code of the header of a web page. Generally however, the information collected is not displayed to the user, but is recorded, tracked, and stored by the user's computer and browser. The website can read the cookie later to identify the personal preferences. Such information will enable the user to navigate the website more easily on return visits. Websites, for instance, can recall registration information, so that users need not re-register each visit. Similarly, cookies enable each user to move forward and backward within a site each session. Most cookies last during a user's "session.". Some can be programmed to last forever -- persistent cookies -- with

the corresponding power to keep track of the user's movements on the Web.

Marketers can then use information about an individual's use of a site to tailor and fine tune sales and promotional offers to consumers, whether on the Web, via e-mail, or at home. Marketers bring information to those who may not know of particular goods and services. Information links sellers to willing buyers, helping achieve a more efficient economy. To some extent, individuals who choose to participate in commercial transactions must give up some personal information to have access to credit and other financial services.

Such information, however, can also be used to reveal all of our personal habits. If marketers share information with each other, an entire mosaic is created revealing our buying patterns, our browsing interests, and the time we spend on the Internet. Many fear the adverse consequences if that information gets into the wrong hands. Estimates suggest that the average American is listed on many computerized databases.

Individuals can disable cookies by setting their browsers not to accept them. Some websites will not do business with such users, and in any event, disabling cookies makes navigation through websites quite cumbersome.

Fifthly, the Internet permits data marketers to pull together a vast amount of information easily. Public records are aggregated on many Internet sites.

Sixthly, companies have programmed "bots" or spiders to canvas the web and retrieve personal information on other sites, usually email addresses. Thus, a third-party can with ease harvest email addresses and other identifying information supplied to a website. Although websites protect financial information through secure socket layer (and other) technology, less sensitive information can be obtained.

10. Is there any Legal Protection for Data Privacy in Sri Lanka?

Information about an individual's tastes and leisure activity has economic value, and the exchange of such information helps to grease the economy. Sri Lanka has never banned the sale of such data, despite the potential impact on privacy. There are, however, many different levels of legal protection for privacy when websites and e-commerce firms, without consent, use private information for commercial purposes. No comprehensive protection exists. The following covers the constitutional & other legal protection for individual privacy in Sri Lanka.

In many countries around the world, there is a general law that governs the collection, use and dissemination of personal information by both the public and private sectors. An oversight body then ensures compliance. This is the preferred model for most countries adopting data

protection laws and was adopted by the EU to ensure compliance with its data protection regime.

11. International Agreements to Protect Privacy

There are three principal international agreements, which are of general relevance to data privacy [4]:

1. The Organization for Economic Cooperation and Development's (OECD) Guidelines on the Protection of Privacy and Trans border Flows of Personal Data of 23 September 1980
2. The Council of Europe Convention No 108 for the Protection of Individuals with regard to the Automatic Processing of personal data adopted 28 January 1981
3. The International Covenant on Civil and Political Rights in 1966 (ICCPR) (and its European equivalent)

Apart from these agreements is the European Union Council Directive 95/46/EC entitled "Directive on the Protection of Individuals with regard to the Processing of Personal Data and the Free Movement of Such Data" [5] which was adopted on October 24, 1995 and the United Nations General Assembly guidelines for the Regulation of Computerized Personal Data files on December 14, 1990. In July 2000, the European Commission, which issued a proposal for a new directive on "the processing of personal data on the protection of privacy in the electronic communications sector". This replaces the 1997 EU Telecommunications Directive and the General Agreement on Trade in Services (GATS) (Stating in art XIV that member states are not prevented by this worldwide agreement to adopt or enforce regulations relating to the protection of privacy of individuals in relation to the processing and dissemination of personal data and the protection of confidentiality of individual records and accounts)

12. Constitutional Protections

The 1978 Constitution of the Democratic Socialist Republic of Sri Lanka does not explicitly recognize the right to personal privacy as a basic fundamental right. In October 1997 and the year 2000 the proposed Constitutions envisaged right to privacy as a fundamental right. The proposed October 1997 Constitution's Article 14 (1) specifically states, "Every person has the right to respect for such person's private and family life, home, correspondence and communications and shall not be subjected to unlawful attacks on such person's honour and reputation. Therefore unlike the U.S.A there is no

reasonable expectation of privacy against intrusions by the state.

13. Legislation

The government has not introduced any specific legislation, which protects the individual privacy or collection of personal information. The only legislation, which refers to this area, is the Telecommunication act No 27 of 1996 and that too refers to interception of communication.

14. The Indirect Protection of Privacy

The Common law in Sri Lanka does not recognize any right to protect personal information. It only permits peripheral protection or remedial action for, invasions of privacy stemming from the inappropriate use of personal data.

15. Contractual Liability

It is possible to include the terms of a contract express protection for personal information. Typically, such provisions are broader than just personal information. They extend to the protection of all information flowing between the parties to the contract. These types of clauses supplement any existing rights the parties which may already have under the tort of breach of confidentiality. The law also implies a number of protections into a variety of contractual relationships. But the contractual relationship is not the essential ingredient, which has given rise to these protections. It is rather the confidential nature of the relationship. Special relationships exist between banks and customers, doctors and patients and lawyers and clients. They may also exist in a non-contractual context; for example, the confidentiality that exists between priests and their parishioners. The ability of the law of contract to provide a solution is severely limited because the data subject is not in a contractual relationship with the data collectors or users. Thus there are no express or implied contractual rights bestowed upon the data subject.

16. Tortious Liability

16.1 Negligence

There are various possibilities in tort. The most obvious possibility would be an action brought by the data subject against the data controller for negligent use of storage of the data. For instance, a third party has gained unauthorized access to personal data about the data subject due to the direct or vicarious negligence of the data controller. Such an action will be possible only

where a duty of care owed by the data controller to the data subject is established, and this will involve inter alia a consideration of the nature of the information.

16.2 Trespass

Trespass consists of the wrongful entry by the defendant onto land belonging to the plaintiff without consent, the plaintiff being the rightful possessor of the land. Where access to personal data is achieved through the unauthorized access to a computer, which is accomplished in turn by the wrongful physical entry of the defendant upon the plaintiff's premises, an action will obviously lie for the trespass to land, though this has been incidental to the main objective of gaining access to personal data.

16.3 Defamation

Defamation is a cause of action intended to protect the reputation of a person whose standing has been lowered in the estimation of "right thinking members of society" by the publication of a derogatory and untrue statements. The electronic dissemination of derogatory statements about a data subject through discussion groups or other Internet facilities will provide the subject matter with a cause of action in defamation provided that they are untrue.

16.4 Intentional Inflicting of Distress

Where a person intentionally or recklessly conducts himself so as to cause emotional distress to others, he is liable for that distress. Conceivably, the same course of action would lie where a person revealed personal information designed to cause the data subject acute embarrassment. It is essential that the injury suffered be of an enduring or physical nature.

16.5 Misfeasance in Public Affairs

This tort requires misconduct by the holder of a public office. Such persons must owe the public duties in the manner in which the administrative duties of the officer are performed. The breach of the statutory duties of confidence imposed upon public servants in relation to personal data accessed or used in the course of their administrative duties, for example, would give an aggrieved data subject a cause of action.

17. A Mechanism for Addressing Data Privacy Issues

Global consistency is fundamental to achieving effective privacy protection. If different standards and approaches are taken, the confusion that would result could well

undermine rather than enhance consumer protection and it could hinder the development of E-business. If one stand is to be adopted globally, I suggest that the Informational Privacy Principles based on OECD guidelines and European Directives would be a practicable solution. Therefore the Sri Lanka draft Data Protection law should be based on these principles. All these principles are based on protection of the individual privacy protection [6].

18. What are Informational Privacy Principles?

Principle One - Manner and purpose of collection of personal information

Personal information must not be collected where it is gathered by unlawful means; for example theft. This principle extends on prohibition to collection by unfair means. It further elaborates that such collection must be necessary for or directly related to that purpose.

Principle Two - Solicitation of personal information from the individual concerned

This principle is designed to ensure that agencies that collect personal information take steps to make the data subject aware of the purpose for which the information is being collected and, where the information is passed on by the collector, the details of the person or persons who receive the information. This principle only applies where the collector solicits the information from the data subject or individual concerned.

Principle Three - Solicitation of personal information generally

Where information is collected through a process of solicitation, the collector must ensure that reasonable steps are taken to determine the relevance, completeness and currency of the data. In addition, this principle requires that the information collected does not unreasonably intrude upon the 'personal affairs' of the data subject.

Principle Four - Storage and security of personal information

This is an important principle, which lies at the heart of the integrity and security of the personal information that is collected and stored. The term 'record keeper' is introduced here. The record-keeper must ensure that security safeguards that are appropriate in the circumstances are taken to prevent loss, unauthorized access, use, modification or disclosure other misuse of information.

Principle Five - Information relating to records kept by a record-keeper

This is the openness principle. It requires a record-keeper to have in place a system to enable data subjects, or any other person, to determine whether a record-keeper has possession or control of any records containing personal information.

Principle Six - Access to records containing personal information

The availability of access by data subjects to material in the possession or control of record-keepers is central to any privacy regime. This principle sets out that the data subject is entitled to have access to these records without excessive delay or expense, except where the record-keeper is required or authorized to refuse access pursuant to any law that provides access by persons to document.

Principle Seven - Alteration of records containing personal information

This principle sets out the rights of the data subject in relation to ensuring the quality of the information held about him or herself. Appropriate corrections, deletions and additions are required to ensure that the record of personal information conforms to the principle.

Principle Eight - The record-keeper to check the accuracy etc. of personal information before use

A record-keeper who has possession or control of a record that contains personal information shall not use that information without taking such steps (if any) as are, in the circumstances, reasonable to ensure that, having regard to the purpose for which the information is proposed to be used, the information is relevant accurate, up to date complete and not misleading.

Principle Nine - Personal information only to be used for relevant purposes

The intention of this provision is clearly to prevent misuse of information where it is not relevant to the purpose for which it will be used.

Principle Ten - Limits of disclosure of personal information

A record-keeper who has possession or control of a record that contains personal information shall not disclose the information to a person, body or agency (other than the individual concerned) unless, the individual concerned has been informed that the information of that kind is usually passed to that person, body or agency or the individual concerned has consented to the disclosure and the disclosure is required or authorized by law.

Principle Eleven - Sensitive information

Any information relating to ethnic or racial origin, political opinions, religious or philosophical beliefs, trade union membership, health or sexual life shall not be used

or disclosed by a record-keeper without the express written consent, freely given, of the individual concerned. Information relating to an individual's criminal history may only be processed as required or authorized by law.

19. The Role of Data Protection Commissioner

An essential aspect of any privacy protection regime is oversight. In most countries with a data protection act, there is also an official or agency that oversees enforcement of the act. This must be absolutely an independent supervisory authority. Independence is also a problem in many countries where the agency is under the control of political arm of the government or part of the ministry and is given considerable power; Government must consult this agency when the government draws up legislation relating to the processing of personal information. The body also has the power to conduct investigations and have a right to access information relevant to their investigations, impose remedies such as ordering the destruction of information or ban processing, and start legal proceedings, hear complaints and issue reports [7]. The agency is also generally responsible for public education and international liaison in data protection and data transfer. It should also maintain the register of data controllers and databases. Another significant feature of this body is that the agency issue guidelines and drafts regarding industry code of conduct and practice for public consultations before it implements.

A major problem with many agencies around the world is a lack of resources to adequately conduct oversight and enforcement. Independence is also a problem. In many countries, the agency is under the control of political arm of the government or part of a particular Ministry and may lack the power or will to advance privacy or criticize privacy invasive proposals.

20 Conclusion

Unlike the European Union, the United States traditionally has adopted a different approach to data protection. The European Union embraces privacy as a fundamental right and thus considers comprehensive legislation as the most appropriate means to protect personal information. Such an approach requires the creation of government data protection agency and approval before the processing of persona data. By contrast, many Americans believe in the free market and are constantly suspicious of government intrusions. The U.S approach relies on a mix of legislation, administrative regulation and industry self-regulation through code of conducts developed by industries as an alternative to government regulation. In my opinion, I firmly believe If Sri Lanka is really willing to accept the benefits of the

globalization and getting absorbed into International trade we still are not too late for any proposed data protection law that should be based on European model of the EU directive and the data privacy principles because U.S Industry self-regulations are more flexible and there is no independent authority to protect and implement data users rights. Finally we should recognize data privacy as one of our fundamental rights. We need more laws in the emerging new area to attract more e-business through out the world.

Bibliography

1. UNCTAD, E-Commerce and Development Report
2. MD Kirby, 'Human Rights and Technology: A New Dilemma' (1988) 22 University of British Columbia Law Review 123 at 127
3. R Wacks, 'Privacy in Cyberspace: Personal Information, Free Speech and the Internet' in P Birks (ed) Privacy and Loyalty Oxford (1997) at 93.
4. HH Perritt and CJ Lhulier, 'Information Access Rights Based on International Human Rights Law' (1997) 45 Buffalo Law Review 899 at 906
5. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal No 281, 23/11/1995 p 31
6. SD Baiz and O Hance, 'Privacy and the Internet: Intrusion, Surveillance and Personal Data' (1996) 10(2) International Review of Law, Computers and Technology 219
7. Brian Foran, Office of the Privacy Commissioner of Canada, 'The Role of the Federal Privacy Commissioner'. Panel Presentation to E-Commerce and Privacy Conference, Ottawa, Ontario, February 21, 2000.
http://www.privcom.gc.ca/english/02_05_000221_2_e.htm



Hybrid Ant Colonies for Parallel Algorithms

W R M U K Wickramasinghe¹, D N Ranasinghe²
 Department of Computation and Intelligent Systems
 University of Colombo School of Computing
 35, Reid Avenue, Colombo 7,
 Sri Lanka.

E-mail: ¹ukw@ucsc.cmb.ac.lk, ²dnr@ucsc.cmb.ac.lk

Abstract

To find solutions to Combinatorial Explosive problems such as NP Complete problems require high computational power. Parallel Algorithms can be used to generate programs which can be used to find solutions for them in finite time. Many approaches are used to develop better parallel algorithms, where Bio-Inspired approaches have proved to be more successful than other approaches

Keywords: Ant Colony Algorithms, Hybrid, Genetic Algorithms, Parallel Algorithms, Speedup, Cluster Computing.

1. Introduction

The solving of NP (Non-deterministic Polynomial-time) Complete problems can be very time consuming even by using the fastest computers to date. These results may be crucial and the speed of the outcome is very important. One good way to achieve this speedup is to use parallel computing.

Making a serial program to run in a parallel computing environment is quite challenging. This is because the main factor behind making the code parallel is that each processor used for computation should be well occupied, so that when some are very busy most of the time others are not left idle. In a given serial program the idea in transforming it to a parallel program is to make as much of it able to run in parallel. Traditional approaches making most part of the code capable of running in parallel are not very adequate. Different other approaches can be used to accomplish this more successfully. This paper presents a combination of the Ant Colony Approach and Genetic Algorithms.

1.1 Aims and Objectives

This paper investigates the Ant Colony Algorithm [4] [5] [6] and the optimization done using Genetic Algorithms [16], and proposes a hybrid approach which, can be used to form algorithms for parallel computing. With this Hybrid Ant Colony and Genetic Algorithm, the main objective is to demonstrate that efficient parallel algorithms can be developed. The case study used to test and demonstrate the prototype is the Traveling Salesman Problem (TSP). The implementation will run on a cluster of homogeneous computers.

The following specific aims have been identified.

- Investigate the literature on Ant Colony algorithms and the Genetic Algorithm optimization of it.
- Develop a Hybrid Ant Colony Genetic Algorithm
- Obtain a Parallel code which will use all processors in the Cluster efficiently.
- Simulate an Ant Colony on a Cluster of Computers.
- Find faster solutions for the TSP.
- Find solutions for city numbers up to 5000.

2. Background

2.1 NP Complete Problems

NP Complete – Non-Deterministic Polynomial time Complete – problems are the type of problems which can be solved but finding a solution is very time consuming. This requires time which cannot be even calculated before hand. These types of problems are called NP Hard problems also. The TSP is one of the most studied NP Complete problems.

2.2 Traveling Salesman Problem (TSP)

The TSP is a situation where, there is a salesman who has to visit a given number of cities. He has to move from one city to another, visiting each city only once and come back to his original starting point. This path of travel should be the shortest path. This being the reference problem used for this dissertation, a general definition to the TSP is given in the following manner.

Consider a set of N nodes, representing cities, and a set E of arcs fully connecting¹ the nodes N . Let d_{ij} be the length of the arc $(i,j) \in E$, that is the distance between cities i and j , with $i,j \in N$. The TSP is to find the minimal length Hamiltonian circuit on the graph $G = (N, E)$, where an Hamiltonian circuit of graph G is a closed tour visiting once and only once all the $n = |N|$ nodes of G , and its length is given by the sum of the lengths of all the arcs which it is composed of [7].

It is needed to be noted that the distances need not be symmetric. In an asymmetric TSP (ATSP) $d_{ij} \neq d_{ji}$. This text will be looking into symmetric TSP situations only.

Looking at the above mentioned NP Complete problem, it is necessary to obtain the results (solutions) as quickly as possible. In the case of real world problems time constraint is very important. The next area will look into how to find solutions to NP Complete problems. In addition the next area would also discuss how the speedup can be achieved using parallel algorithms.

3. Hybrid Ant Colony-Genetic Algorithm Approach

The proposed algorithm is a Hybrid approach of Ant Colony Algorithm and optimizing it by Genetic Algorithms.

Ant Colony Algorithms was first introduced by Marco Dorigo [4] [5] [6]. Today it is being widely used in many situations like solving NP Complete problems to load balancing in telecommunication networks [11]. How to use Ant Colony Algorithms and optimize it by Genetic Algorithms to achieve a better parallel algorithm to solve TSP? First it is interesting to see how real ants behave.

3.1 Real Ants

One of the most interesting scenes in nature is to see ants traveling from their nest to a food source and

back, all in one line. This is more intriguing given the fact that ants are almost blind animals with a very little intelligence. How do they manage to find the shortest path from their colony to the food source and back? It was found that the medium they use to communicate with each other about paths they took and should travel, had trails of pheromone². A moving ant lays some pheromone (in varying quantities) on the ground, marking a path by a trail of this substance. An isolated ant moves in random, an ant encountering a previously laid trail can detect it and decide with high probability to follow it, which will reinforce the trail with its own pheromone. These collective behaviors of ants will determine a route for others to follow. The probability of with which an ant chooses a path increases with the number of ants that previously chose the same path.

What will the ants do, if they encounter an obstacle on their path along which they are happily moving? In the following diagram it is seen that in the detection of an obstacle ants in random move around it, all the while marking the path with their pheromone. The pheromone is a substance which evaporates. When ants go on a shorter path, the pheromone trail is being updated quickly than the longer routes. Thus the pheromone density on a shorter route will be higher than of the longer routes. The ants following will move along the route where the pheromone trail is stronger. This way all the ants will take the shortest path to avoid the obstacle.

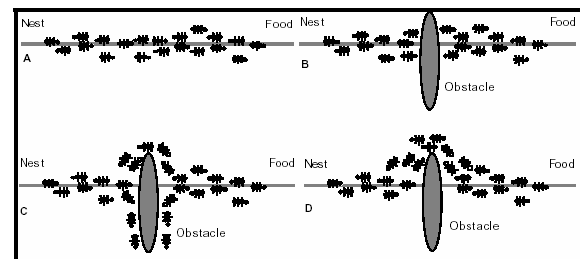


Figure 3-1: Ants moving along a path with an obstacle (A) Ants follow a path between the nest and food sources. (B) An obstacle appears on the path: Ants choose whether to turn left or right with equal probability. (C) Pheromone is updated more quickly on the shortest path. (D) All ants have chosen the shortest path.

This way all ants are capable of moving from one location to another on the shortest path, avoiding any obstacle in their way [4] [5].

¹ Fully Connected Graph: A graph where there exists a route from any node i to node j , where $i \neq j$.

² Pheromone: substance secreted and released by an animal for detection and response by another, usually of the same species.

3.2 Ant Colony Algorithm

With some small modifications to “Real Ants”, artificial ants are defined for the solving of the TSP. This section describes how these modified ants can be used to obtain solutions for the TSP.

The Ants used here are assumed to have a small memory of the cities it has visited and can perform calculations to determine a city to move to. Let there be a symmetric TSP with n cities. The number of ants is m , which is constant over time. For an ant located at city i , the transition from i to j depends on,

- whether or not city j has been visited. Ants have that information in their memory M_k for the k^{th} ant.
- the distance d_{ij} between i and j . $d_{ij} = d_{ji}$.
- the amount of pheromone on the edge connecting i to j , denoted by $\tau(i, j)$.

The ant will move from i to j according to the following probabilistic formula,

$$s = \begin{cases} \arg \text{Max}_{j \notin M_k} \{ \tau(i, j) [\eta(i, j)]^\beta \} & \text{if } q \leq q_0, \\ S & \text{otherwise} \end{cases}$$

Here,

$\eta(i, j)$ is the inverse of d_{ij} .

$\beta \geq 0$ is a parameter for relative importance of pheromone trail

q is a value chosen randomly with uniform probability in $[0, 1]$

$0 \leq q_0 \leq 1$ is a parameter

S is a random variable selected according to the following probability distribution, which favors edges which are shorter and have a higher level of pheromone trail:

$$p_k(i, j) = \begin{cases} \frac{[\tau(i, j) [\eta(i, j)]^\beta]}{\sum_{j \notin M_k} [\tau(i, j) [\eta(i, j)]^\beta]} & \text{if } j \notin M_k \\ 0 & \text{otherwise} \end{cases}$$

Here $p_k(i, j)$ is the probability with which ant k chooses to move from city i to j .

The pheromone trails is then updated both *locally* and *globally*.

Local update formula: This is done to avoid a very strong edge being chosen by all ants, which is not the

shortest. This update is motivated by the evaporation of pheromones. (p, q) is a visited edge connecting city p and q .

$$\tau(p, q) \leftarrow (1 - \alpha) \cdot \tau(p, q) + \alpha \tau_0$$

Where $\alpha \geq 0$ is the relative importance of a trail $\tau_0 > 0$ is a parameter.

Global update formula: This intends to reward the edges belonging to shorter tours. (p, q) is a visited edge.

$$\varphi(p, q) \leftarrow (1 - \alpha) \cdot \varphi(p, q) + \alpha \cdot \Delta \varphi(p, q)$$

Where $\varphi(p, q)$ is the same as $\tau(p, q)$. $\Delta \varphi(p, q)$ is the inverse of the value of the shortest tour [2] [4].

These are the basic functions used in the ant algorithm. Next section will investigate how a parallel algorithm can be obtained by using this ant algorithm optimized by a genetic algorithm.

3.3 Proposed Approach (Hybrid)

Since this is a combined approach of both Ant and Genetic Algorithms, it is referred to as a Hybrid approach. The term Hybrid Algorithm will be used from here onwards for the proposed algorithm.

This Hybrid Algorithm depends on several parameters; they are the number of generations which the algorithm is to run, initialization of the pheromone trail, pheromone updating and pheromone evaporations. This is done using local and global update formulae [14] [16].

Hybrid Ant Colony-Genetic Algorithm

```

Begin
  Initialize AntColony;
  for n=1 to NoOfGenerations do
    Begin
      For k=1 to NoOfAnts do
        Begin
          Locate Ant(k) in starting positions at random;
          Use Ant(k) to solve TSP;
        End;
      Perform local update on pheromone trails;
      Obtain the shortest path by Ant(k);
      Perform global update on pheromone trail;
      Update parameters for Generation(n+1) on
        Generation(n);
    End;
  Project shortest path and other results;
End.

```

Figure 3-2: Proposed Hybrid Ant Colony-Genetic Algorithm

This proposed algorithm is shown to be one which can be used as a parallel algorithm. This can be achieved in the following manner.

3.4 How Parallelism is obtained

After locating each Ant at a starting point they work on their own traveling through the given city map. This scenario can be very well modeled as the Diffusion Model³ [18]. At the end of each generation the local pheromone update is done. After that the global update is done from the shortest tour made by an Ant. Here each of the Ants' pheromone trails are updated using the shortest toured Ant's traversed distance. At this point the Generation's parameters are also updated using this value. This execution of steps acts like the Island Model⁴ [18].

Therefore this algorithm itself works according to two parallel algorithm models, which makes it a parallel algorithm. The next section will depict how this is implemented in a parallel computer environment, with detailed workings of the algorithm.

4. Implementation

The prototype implementation was done on a cluster of homogeneous computers at the University of Colombo School of Computing. The following are the hardware specifications.

- Cluster with 4 Nodes
- Each node having Dual PII 550 MHz processors
- Each having 256 Mb of RAM
- Connected with Ethernet (100 Mbps)

The operating system running on the cluster is Red Hat Linux 7.2 [9]. Each node has an identical copy of the operating system running. The standard input and output is handled by the root node.

The implementation was done using C language.

Message passing for the distributed memory architecture was achieved using MPI (Message Passing Interface). The version of the MPI library used here was MPICH 1.2.4 [8].

The cluster simulates an Ant Colony, where each node simulates one Ant. Each node is ranked from 0

to 3, so that Node0 is the root node. The following are the steps detailing the algorithm.

Initialize the Environment

- a. Read the city map from Node0 and broadcast the data to all nodes. The location of a city is given by 2 dimensional co-ordinates.
- b. Each node initializes the Pheromone Matrix using the Empirical values from a file. This Pheromone Matrix will be used to simulate the pheromone trails of the ants.
- c. Initialize the time on each node, starting the execution.

Execution

- d. The execution is done for several iterations, where each iteration is called a generation.
- e. Solve the given TSP in parallel at each node. This simulates ants traveling the given city and arriving at the starting point. All the parameters used within the solution finding process are Empirical values read from the header file previously.
- f. Perform local updates of the pheromone trail in each node independently. This simulates the pheromone evaporation on the trails of the ant.
- g. Perform global updates of the pheromone trail taking to account the distance traveled by each ant in each node independently. This simulates the strengthening of pheromone trails of shorter distances, so longer paths are avoided.

Genetic Algorithm Optimization

- h. The parameters used for the solving of TSP are updated considering the shortest path traveled so far by an ant at the end of the iteration.
- i. The new updated pheromone trail is then broadcast to each node. This simulates that all ants are traveling on the same environment. This is done to simulate a shared memory environment on a distributed memory system. So when looking from outside one sees that all are traveling in the same environment.
- j. The best-fit values are used to create the next iteration, so that better solutions can be achieved.
- k. The path which gives the best tour is recorded. This is overwritten when better tours are found during successive generations.

³ Diffusion Model: This is an extreme case of the Island Model, where each process works independently and produces a result at the end of the execution.

⁴ Island Model: All processes work independently, but they exchange their computation values at the end of a given round, according to a rule. Then continue working on those results.

Producing Results

- l. Stop the timer and get the average time taken to make one tour considering the worst time values.
- m. Output the best tour distance traveled by an ant at the end of the execution.
- n. Output the path of the best tour.

Since the cluster simulates the ant colony, each ant's behavior is done in parallel. This approach makes the parallel executing region very significant but message passing between each processor is also kept to a minimum. All these are set to achieve a significant speedup.

5. Results

5.1 Speedup

Here time taken to generate 1000 ant cycles (or ant tours) is given for five datasets. They are given as time taken by using 1 processor and by 4 processors run in parallel. All times are given in seconds.

Problem Name	Time taken on 1 Processor	Time on 4 Processors in parallel	Speedup
Eil51	6.20	2.20	2.81
Eil76	13.81	4.12	3.35
KroA100	24.13	6.72	3.59
U574	838.28	211.11	3.97
Fl1557	6258.70	1567.23	3.99

Table 5-1: Speedup Achievement

Consider the average speedup as 3.5 (assume that the inter-process communication time has been neglected for calculating the speedup) and by applying the Amdahl's Law it's seen perceived that 95% of the code can be run in parallel. So this means only 5% of the code cannot be run in parallel.

According to the Gustafson's Law [17] the scaled speedup can be given as,

$$S_s(n) = n + (1 - n) s$$

$$4 + (1 - 4)0.05 = 4 - 0.15 = 3.85$$

5.2 Results for datasets

Problem Name	Best Solution	Optimum Solution
Ulysses16	74	74
Eil51	426.21	425
Eil76	437.75	435
KroA100	21,427.95	21,282
U574	37,386.71	36,905
Fl1577	22,447.51	22,249
U2319	245,100.31	234,256
Pcb3038	138,819	137,694
Fn14461	184,670.42	182,566

Table 5-2: Results for the datasets used in the testing

The following parameter values were set in order to obtain the above mentioned results.

$\beta = 1.5$ - Relative importance of pheromone trail and of closeness, $\beta \geq 0$

$\rho = 0.9$ - Pheromone persistence (1 - ρ) is the evaporation rate of pheromone, $0 \leq \rho < 1$

$\alpha = 1.5$ - Relative importance of the trail, $\alpha \geq 0$

$\tau_0 = 1.2$ - Parameter: Such that $\tau_0 > 0$

$q_0 = 0.1$ - Parameter: Such that $0 \leq q_0 \leq 1$

5.3 Comparison with other approaches

The results of the proposed Hybrid Ant Colony Genetic Algorithm (HACGA) have been compared with several other approaches. These are compared with the same data sets with approaches such as Ant Colony (AC), Genetic Algorithm (GA), Simulated Annealing (SA), Evolutionary Programming (EP) and Annealing Genetic Algorithm (AG). The compared results were extracted from "Ant Colonies for the TSP" [5] research paper.

Probl em	HACGA	AC	GA	SA	EP	AG
Eil51	426.21	425	428	443	426	436
Eil76	537.75	535	545	580	542	561
KroA 100	21,427,59	21,282	21,761	N/A	N/A	N/A

Table 5-3: Comparison with other approaches

The Hybrid Algorithm performed better most of the different approaches separately. Though it didn't obtain the optimum value of the dataset, the best distance achieved was close to the optimum value. The timings for the other approaches were not available in the referred paper [5], even though comparing the times would be unfruitful since the

hardware in which the Hybrid Algorithm was tested differs from the referenced research paper.

6. Conclusion and Future Work

The Hybrid Approach proved to be better than using a single approach to solve a NP Complete problem like TSP. In some cases the solutions converged to the optimum value and some were very close to the optimum solutions. A very interesting observation identified was that time taken to find solutions were finite and minute. This shows that solutions are found quickly, proving the efficiency efficacy of the code.

The main interesting point observed in the analysis was the speedup. On average the speedup for a 4 processor cluster was noted as 3.5. This had a pattern showing as the size of the dataset increases the speedup increased as well. On some occasions the speedup was close to 4. So the Hybrid Algorithm has proved to be very efficient as a parallel algorithm for large TSPs.

Several areas of improvements or extensions to this research can be proposed. One is the use of other combined approaches (or Hybrid Approaches) for the Ant Colony to better solutions.

Another extension can be done in the area of message passing. This can be a combination of message passing and threading for shared memory architectures. It would be an interesting approach for better parallel codes that can be used on the given hardware configuration of the cluster of computers.

Acknowledgements:

I would like to thank my supervisors Dr. Nalin Ranasinghe and Dr. Prasad Wimalaratne for providing me with many reference material and guidance to make this research a success. I thank Mr. Malik Silva, who helped me a lot in debugging my C program and introducing me to message passing on the cluster environment. Many thanks go to Mr. Ziyan Marikkar for his help when I faced problems in the cluster of computers and also for providing up-to-date MPI libraries. Mr. Damon Cook of the New Mexico State University helped me in many ways via E-Mail to gather resources to compile this research. I thank him very much for his help which was very useful in many ways. Last but not least I'm grateful to Mrs. Himadhu Kottege for proof reading this text. Thank you all.

References:

- [1] Bartak, R. 1998, *Heuristic and Stochastic Algorithms*, Available at: <http://kti.ms.mff.cuni.cz/~bartak/constraints/stochastic.html>
- [2] Botee, H. & Bonabeau, E. 1998, "Evolving Ant Colony Optimization", *IEEE Advance Complex Systems*, pp. 149-159.
- [3] Cook, D. 2000, *Optimizing the Parameters of an Ant System Approach to the Traveling Salesman Problem Using a Genetic Algorithm*, New Mexico State University. (B.Sc. Dissertation)
- [4] Dorigo, M., Maniezzo, V. & Coloni, A. 1996, "The Ant System: Optimization by a colony of Cooperative Agents", *IEEE Transaction on Systems, Man and Cybernetics*, Part-B, Vol.26, No.1, pp.1-13.
- [5] Dorigo, M. & Gambardella, L. M. 1997, "Ant Colonies for the Traveling Salesman Problem", *BioSystems*.
- [6] Dorigo, M. & Gambardella, L. M. 1997, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem", *IEEE Transactions on Evolutionary Computations*, Vol.1, No.1.
- [7] Dorigo, M., Di Caro, G. & Gambardella, L. M. 1999, "Ant Algorithms for Discrete Optimization", *Artificial Life*, Vol 5, No.3 pp. 137-172.
- [8] MPICH, Available at: <http://www.mcs.anl.gov/mpi/mpich/index.html>
- [9] Red Hat Linux, Available at: <http://www.redhat.com>
- [10] Robbe, N. *A Comparison of Optimizing Techniques*, Available at: <http://www.kbe.co.za/usergrp/1998/llog/robbe.htm>
- [11] Schoonderwoerd, R., Holland, O., Bruten, J. & Rothkrantz, L. "Ant-based load balancing in telecommunications networks".
- [12] Stutzle, T. & Dorigo, M. 1999, "ACO Algorithms for the Quadratic Assignment Problem", *New Ideas in Optimization*, McGraw-Hill.
- [13] Stutzle, T. & Dorigo, M. 1999, "ACO Algorithms for the Traveling Salesman Problem", *Evolutionary Algorithms in Engineering and Computer Science: Recent Advances in Genetic Algorithms, Evolution Strategies, Evolutionary Programming, Genetic Programming and Industrial Applications*, John Wiley & Sons.
- [14] Talbi, E., Roux, O., Fonlupt, C. & Robillard D. "Parallel Ant Colonies for Combinatorial Optimization Problems".

- [15] *TSPLIB*, Available at:
<http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/>
- [16] White, T., Pagurek, B., Oppacher, F., “ASGA: Improving the Ant System by Integration with Genetic Algorithms”, *Systems and Computer Engineering*, Carleton University.
- [17] Wilkinson, B. & Allen, M. 1999, *Parallel Programming*, Prentice Hall Press, ISBN 0-13 671710-1
- [18] Zomaya, A. Y., Ercal, F. & Olarian, S. 2001, *Solutions to Parallel and Distributed Computing Problems*, Wiley Interscience Publication, ISBN 0-471-35352-3



A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation

Ruvan Weerasinghe
Department of Computation and Intelligent Systems
University of Colombo School of Computing
Colombo, Sri Lanka.
arw@ucsc.cmb.ac.lk

Abstract

Data-driven approaches to Machine Translation have come to the fore of Language Processing Research over the past decade. The relative success in terms of robustness of Example Based and Statistical approaches have given rise to a new optimism and an exploration of other data-driven approaches such as Maximum Entropy language modeling. Much of the work in the literature however, largely report on translation between languages within the European Family of languages. This research is an attempt to cross this language family divide in order to compare the performance of these techniques on Asian languages. In particular, this work reports on Statistical Machine Translation experiments carried out between language pairs of the three major languages of Sri Lanka: Sinhala, Tamil and English. Results indicate that current models perform significantly better for the Sinhala-Tamil pair than the English-Sinhala pair. This in turn appears to confirm the assertion that these techniques work better for languages that are not too distantly related to each other.

Keywords: Statistical Machine Translation, Language Modeling, Sinhala Language Processing, Tamil Language Processing.

1. Introduction

Machine processing of Natural (Human) Languages has a long tradition, benefiting from decades of manual and semi-automatic analysis by linguists, sociologists, psychologists and computer scientists among others. This cumulative effort has seen fruit in recent years in the form of publicly available online resources ranging from dictionaries to complete machine translation systems. The languages benefiting from such exhaustive treatment however tend to be restricted to the European Family, most notably English and French.

More recently, however, the feasibility of data driven approaches in the context of today's computing power holds out hope for the rest of us. These are those of us concerned with less studied languages, who have few or no linguistic resources to assist us, and for whom the cost of building these up from scratch is prohibitive.

1.1 Background

Sinhala is a language of around 13 million people living in Sri Lanka. It is not spoken in any other country, except of course by enclaves of migrants. Being a descendant of a spoken form (Pali) of the root Indic language, Sanskrit, it can be argued that it belongs to the large family of Indo-Aryan languages.

Tamil, the second most spoken language in Sri Lanka with about 15% of its 18 million people counting as native speakers, is also spoken by some 60 million Indian Tamils in Southern India. The dialects of Indian Tamil however differ significantly to those of Sri Lankan Tamil to cause difficulty in performing certain linguistics tasks [1]. Though originating in India, it does not share the Indo-Aryan heritage of Sinhala, but rather is one of the two main languages in another (unrelated) family known as Dravidian.

Over many centuries of co-existence especially within Sri Lanka, these two languages have come to share many of their structures and functions, thus causing them to be closer to each other than their family origins indicate. English, the link language of Sri Lanka, on the other hand, has had a relatively short co-existence with the other two languages, and though also belonging to the large Indo-European language family, has much less in common with Sinhala. Within the context of the past two decades of ethnic strife between the two primary lingua-cultures within Sri Lanka, any prospect that holds some promise of better understanding of each others language becomes a worthwhile task to aim at.

Towards this aim, this research is an exploration of the feasibility of a non-knowledge intensive approach to machine translation.

1.2 The Three Languages: Sinhala, Tamil and English

By far the most studied languages in terms of descriptive as well as computational models of linguistics have been the European ones – among them English receiving more attention than any other. From lexical resources, part-of-speech taggers, parsers and text alignment work done using these languages, fairly effective example-based and statistical machine translation efforts have also been made.

In contrast, electronic forms of Sinhala became only possible with the advent of word processing software less than 2 decades ago. Even so, being ahead of standardizing efforts, some of these tools became so widely used, that almost 2 years after the adoption of the ISO and Unicode standard for Sinhala (code page 0D80), there is still no effort to conform to it. The result is that any effort in collecting and collating electronic resources is severely hampered by a set of conflicting, often ill-defined, non-standard character encodings.

The e-readiness of the Tamil language lies somewhere in between the two extremes of those of English and Sinhala. Here the problem seems to be more the number of different standardization bodies involved, resulting in a plethora of different standards. TASCII was one of the earliest attempts and is an 8-bit Tamil encoding, while the newer Unicode standard is at code page 0B80. Owing to this, available electronic resources may be encoded in unpredictable ways, but tools for translating between non-standard encodings and the standard(s) exist.

1.3 Scope

This work builds on Weerasinghe [6] which explored the application of Statistical Machine Translation between Sinhala and English by proposing a bootstrapping approach to build up the relevant resources. The results reported therein not being too promising, this research is an attempt to find out if the reasons are to do with Asian language characteristics or the ‘linguistic distance’ between the language pair chosen to be translated.

Towards this end, in this work we have undertaken the task of Statistical Machine Translation between Sinhala and Tamil using the same Sinhala corpus and its translated Tamil version for the learning process. The models thus built would

then permit a fair comparison in order to determine the similarity between English-Sinhala translation and Sinhala-Tamil translation.

2. Why Statistical Machine Translation?

Machine Translation as one of the major sub-disciplines Natural Language Processing and Computational Linguistics has as long a tradition as does its parent discipline. Much of the work in the field used a fundamentally knowledge-based approach to the problem, just as they did with Language Processing in general. It was becoming clear by this time that existing (knowledge-based) systems could only prove useful for ‘toy’ worlds and artificially controlled language constructs. Their scalability to the analysis or translation of naturally occurring speech or text was in serious question.

The mid 1980’s saw somewhat of a resurgence in the use of data-driven and statistical approaches to Language Processing in general reflecting its increasing importance in the whole area of Artificial Intelligence. This could be attributed to its effectiveness in learning and adaptability over traditional knowledge-based approaches. By the early 1990’s this revolution had also lead to new efforts into casting the Machine Translation problem also in terms of the data-driven paradigm.

While Example-Based Machine Translation (EBMT) is an example of this approach as applied to the area, the dominant theoretical foundation for such work derived from statistics and information theoretic ideas and came to be known as Statistical Machine Translation (SMT) in general. The most influential work in this regard came from work at IBM’s research labs by Brown et.al. [2].

Knowledge-based approaches to Language Processing, in common with other AI-related tasks, have never been able to satisfactorily solve the knowledge acquisition bottleneck. In Language Processing this translates to spending much time and effort on agreeing on linguistic niceties in order to produce the large linguistic resources required by such an approach.

Data-driven approaches on the other hand require only fundamentally un-pre-processed (‘raw’ forms of) input such as corpora of naturally occurring text from which machine learning can be performed. This makes Statistical Machine Translation an attractive alternative for translation between language pairs for which little or no electronic linguistic resources exist. Sinhala and Tamil are among the many world languages for which this is true.

2.1 The basic SMT model

Statistical Machine Translation is founded upon the assumptions of the Noisy Channel Model and Bayes Rule which help ‘decompose’ the complex probabilistic model that needs to be built for estimating the probability of a sentence in a source language (f) being translated into a particular target language sentence (e).

Using the notation common in the literature this decomposition can be stated as:

$$P(e|f) = P(e) * P(f|e) / P(f)$$

Since predicting in a statistical model corresponds to identifying the most likely translation, maximizing the above over all possible target sentences (e) gives the estimation:

$$\text{argmax}_e P(e|f) = \text{argmax}_e P(e) * P(f|e)$$

The main benefit gained by the above decomposition is that the burden of accuracy is moved away from the single probability distribution $P(e|f)$ to two independent probabilities $P(e)$ and $P(f|e)$. The former is known as the ‘language model’ (for language e) while the latter is known as the ‘translation model’ (for predicting source sentences, f, from target sentences e).

While it would be impossible to estimate such a language model, the literature on using n-gram (mainly bi-gram and tri-gram) models for estimating sentence probabilities of a given language have matured over the past two decades. The estimation of the translation model would not be too difficult if machine readable dictionaries with frequency statistics were available. While this is impractical for even the most well studied languages, the dependence of such counts on the genre of the texts under consideration make it less than optimal.

This is where work carried out by Brown et. al. [2] at IBM stepped into providing a bootstrapping model building process. Beginning with the very simple word-for-word translation lexicon building models (IBM Models 1 and 2), this process constructs ever more sophisticated Models (3, 4 and 5) which account for more and more flexibility in the underlying assumptions (e.g. a single word in the source language may be translated by more than a single target word, and may appear in another part of the sentence).

Intuitively, once the translation model performs its task of predicting a set of possible (good and bad) candidate translations for a particular source sentence, the (target) language model will calculate the probability of such sentences being acceptable in the language in order to select the best translation. It is this ‘sharing of the burden of accuracy’ between

the two models that has been at the heart of the relative success of the SMT approach.

2.2 The SMT process

The SMT process at its very heart requires the compilation of a bi-lingual corpus. The corpus in addition needs to be ‘sentence aligned’: each sentence in the target language must have an identified equivalent source language sentence to which it must be aligned in some way. While this process can be performed manually current research has promising results on automating this process.

The complete SMT process involves (a) the building of a target language model, (b) the construction of the translation model, (c) the decoding process and (d) the process of scoring resultant translations.

While a bi-lingual parallel corpus is the primary and only resource that is needed for applying the SMT process to the language pair concerned, there is no theoretical necessity for the (target) language model, $P(e)$ to be constructed just from its portion in the bi-lingual corpus. It is common practice to augment this with an expanded target language model in order to improve the overall model.

The CMU-Cambridge Toolkit[†] [4] can be used to build such a target language model from a minimally tagged monolingual plain text corpus based on n-gram statistics.

The second component of the SMT process is the building of the translation model as outlined in section 2.1. The IBM models described have been more recently improved in terms of efficiency and made available in the public domain by Al-Onaizan et.al. [3]. This GIZA system provides a set of tools that facilitate the building of translation models from scratch[‡].

Once the translation and language models have been constructed from the training data, the combined model needs to be applied to new test data in order to determine the outcome of the translation process. Since the maximization concerned requires exploring an entire word trellis, a decoding approach is required to determine the highest scoring sentence hypothesis. For this process we used the publicly available ISI-Rewrite decoder with various different parameters and smoothing methods in order to arrive at the best possible scheme for the respective language pair.

[†] Downloadable from svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[‡] Downloadable from www.clsp.jhu.edu/ws99/

While this completes the entire SMT process, the evaluation of the output produced by the system requires a metric that can be applied to any given language pair in order to be able to compare results of different approaches. While many such metrics have been used, they have mostly been based on human judgment and thus not reproducible exactly.

Papineni et. al. [5] suggested a completely automatic metric they referred to as BLUE which is based on the Word Error Rate (WER) metric popular in Speech Recognition. This metric scores between 0 and 1 for any potential translation of a sentence by comparing it to (possibly multiple) professionally translated ‘reference translations’.

3. Sinhala – Tamil SMT

While not requiring large hand-crafted linguistic resources agreed on by linguistic experts, SMT does require a reasonably large parallel bi-lingual corpus whose translations are fairly faithful to its purpose. So, for instance, many obvious sources turn out not to be good candidates for the compilation of such a corpus.

In the Sinhala-English case, as reported in [6], it was found that a major newspaper of Sri Lanka publishing in both languages has two different reporters for each event who report news in very different ways. Even in the context of articles translated from one to the other, it has been found that in most cases, translators have been given full freedom to ‘interpret’ such articles afresh and so end up with translations that have quite different sentence and even paragraph structures. All such candidate sources turn out to be unsuitable for the SMT task. Weerasinghe [6] identified a website (www.wsws.org) as a possible source for the compilation of a Sinhala-English parallel corpus and we here discovered a superset of this corpus as a suitable candidate for a trilingual Sinhala-Tamil-English parallel corpus.

A set of WSWs articles available on the site during 2002, containing translations of English articles into Sinhala and Tamil, were selected to form a small tri-lingual parallel corpus for this research. This consists of news items and articles related to politics and culture in Sri Lanka.

3.1 Basic Processing

The fundamental task of sentence boundary detection was performed employing a semi-automatic approach. In this scheme, a basic heuristic was first applied to identify sentence boundaries and those situations that were exceptions to the heuristic

identified. These were then simply added to an ‘exceptions list’ and the process repeated. This process proved adequate to provide accurate sentence boundary detection for the Sinhala and Tamil corpora.

Automatic sentence alignment proved to be unsuccessful as reported in [6] and hence a manual alignment of sentences was carried out.

After cleaning up the texts and manual alignment, a total of 4064 sentences of Sinhala and Tamil were marked up in accordance with TEI-Lite guidelines. This amounted to a Sinhala corpus of 65k words and a parallel Tamil corpus of 46k words.

3.2 Language Modeling

Owing to the lack of lemmatizers, taggers etc. for Sinhala and Tamil, all language processing done used raw words and were based on statistical information gleaned from the respective ‘half’ of the bi-lingual corpus. The CMU-Cambridge Statistical Language Modeling Toolkit (version 2) was used to build n-gram language models using both the Sinhala and Tamil ‘halves’ of the corpus independently.

Table 1 shows some statistics of the resulting language models with respect to a small test corpus extracted of new articles on the WSWs site. The perplexity figure for Sinhala and Tamil is higher than for English as reported in [6]. However, in both cases larger test sets produced higher percentages of out of vocabulary (unknown) words indicating that the basic corpus size needs enhancing.

For the purpose of building better language models needed for the statistical translation process a monolingual Sinhala (or Tamil as the case may be) corpus needs to be extracted from the same domain.

Description	Sinhala Corpus	Tamil Corpus
Testset	5479 words	5304 words
Perplexity	865.61 (9.76 bits)	1218.86 (10.25 bits)
# 3-grams	327 (5.97%)	258 (4.86%)
# 2-grams	1419 (25.9%)	1227 (23.13%)
# 1-grams	3733 (68.13%)	3819 (72%)
# unseen words	1061 (16.22%)	1247 (19.04%)

Table 1. Perplexities and other statistics for the Sinhala and Tamil WSWs corpora

3.3 Translation model

As discussed in section 3.1, the Sinhala and Tamil bitext was first sentence aligned using semi-automatic means. The resultant parallel corpus contained 4064 sentence pairs in the two languages. In order to construct the Sinhala-Tamil translation model we first used the public domain GIZA toolkit and then also experimented with the newer GIZA++ system of Och et. al. [7].

Many different training schemes were experimented with including the use of a HMM step in place of IBM model 2 using GIZA++. Many different smoothing techniques were also applied in order to arrive at an optimal translation model.

3.4 How Using the SMT model

The ISI-Rewrite decoder was used to generate translations for a set of 162 Tamil test sentences taken from new news items appearing in the WSWWS site. Rewrite was used to select the best 3 translations for visual evaluation purposes. In order to test the system however, only the topmost translation returned was used.

Table 2 shows an example fragment of this text and its translated Sinhala equivalent. As can be seen, the intelligibility itself is affected by the unseen Tamil words which are simply translated by transliteration. What is however more interesting is the evaluation of the these translations against the human translation retrieved from the WSWWS website.

Source Tamil Text
2001 @pbr v`Y 22 @kÄNm v`m @pr Ê Ð a `Nv pl v` h#Pmt vJp`û i ØYpW í n #ñ n Ê W, l < en x`Nw m#ñvr n @Y Û w^n l L h` @J 0 sAi [nv l a Wsì kr g#l m s>h` o 0 0y vr @gn ñ@B.
Target Sinhala Translation
22 පෙබරවාරි 2001 වාච්පායිට් ආන්ව ඛට්ඨර බෙංගාලයේ රිවිජිබ්හුඵ දැමීමට බලයේ සිට වම්විමුල්ප්ඕරිද කරුවන් සිබ්බුබ්බුප්බුද සෝලිය, ප්‍රාන්ත ට්‍රිබිවිබ්හුඵ සහ බෙංගාලයේ විය. පලාතේ සහ සිප්පිය(විම්) පස්දෙනෙක් සිප්පිය(විම්) ප්‍ර ට්ඨිද් ශක්තිමත් ක්‍රියාකාරී තන්ත්‍රයේ ලිඛිල්

Table 2. An example translation of a Tamil test sentence into Sinhala

4. Results & evaluation

In order to carry out an unbiased evaluation of the accuracy of the Sinhala-Tamil translations achieved with respect to the state-of-the-art, we used an implementation of the IBM BLEU score generator.

After testing with multiple translation models, we achieved a best BLEU score of 0.1362 for this task. A straight comparison of this result with the work reported in [6] gives cause for hope. In that work, the best BLEU score achieved for English-Sinhala translations was just 0.0618.

Papineni [5] shows that the BLEU score is sensitive to the number of (expert) reference translations. The most common numbers of reference translations for which scores are quoted in the literature are 2 and 4.

In this work, we had access only to a single translation which was taken to be the reference translation. From [5] it can be gauged that BLEU scores of translations compared with 4 reference translations can be 1.35 times as high as those with 2 reference translations.

Assuming a similar ratio in score differentials between 2 and 1 reference translation(s), the above scores correspond to a Sinhala-Tamil translation BLEU score of 0.185 and a English-Sinhala translation BLEU score of 0.084 with 2 reference translations.

These scores can be compared with the machine translation scores reported in [5] of 0.0527, 0.0829 and 0.0930 and the human translation scores of 0.1934 and 0.2571 – all on 2 reference translations.

Among the obstacles remaining between the current system and more intelligible output translations include (a) the limited size of the corpora highlighted by the high perplexity of the Sinhala and Tamil language models, and (b) the long-distance ‘movement’ of mutually translated words and phrases not captured in current translation models.

In order to address (a), current efforts are underway to extract a larger Sinhala-Tamil parallel corpus as well as larger mono-lingual corpora in each of the two languages concerned. In order to address (b), serious consideration would need to be given to the underlying assumptions of the IBM models and other data-driven techniques pursued where appropriate.

Further work is also planned to combine the information obtainable from all three languages Sinhala, Tamil and English in order to arrive at more accurate models between any two of them.

5. Conclusions and further work

It is apparent from the above results that further work is needed in SMT as a whole to produce intelligible translations. One of the limitations in this particular application of the process is the size of the parallel corpus used for learning.

The dimensions of the Sinhala-Tamil corpus used however being very similar to that of the Sinhala-English corpus reported on in [6], some comparison of the two experiments is warranted.

It is clear from the perplexities of both the Sinhala and Tamil corpora used that their language models are deficient. Despite this however, the Sinhala-Tamil SMT process consistently produced BLEU scores significantly higher than those for English-Sinhala translation reported in [6]. In fact, in most cases, the former score was about twice that of the latter. Closer examination of the type of errors generated in the English-Sinhala case, suggest that sentence structure accounts for many of the incorrect translations. On the contrary, in the Sinhala-Tamil case, sentence structures are much more predictable from each other. This may offer a clue as to the reason for the better performance of the Sinhala-Tamil case.

Further, the above results also point to a possible link between the relative success of SMT for linguistically related language pairs such as English and French as reported in the SMT literature. As such the results of this work contribute to the general body of SMT work by suggesting that SMT between linguistically closely related language pairs perform significantly better than that between linguistically less related language pairs.

This research also provides some reasons for optimism of the general SMT approach for solving the translation problem among Sri Lanka's two native languages, Sinhala and Tamil.

Acknowledgements:

The author wishes to express his gratitude to the US Fulbright Commission for providing a grant to pursue the above work at the Language Technology Institute of the Carnegie-Mellon University, Pittsburgh, USA during 2002. He is also grateful to the University of Colombo for releasing him during this period from his regular duties.

References:

- [1] Germann, U. 2001. Building a Statistical Machine Translation System from Scratch: How Much Bang Can We Expect for the Buck? *Proceedings of the Data-Driven MT Workshop of ACL-01*. Toulouse, France (2001)
- [2] Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J. and Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2) (1993) 263-311.
- [3] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D.: *Statistical Machine Translation, Final Report*, JHU Workshop 1999. Technical Report, CLSP/JHU (1999)
- [4] Clarkson, P.R. and Rosenfield, R.: Statistical Language Modeling using the CMU-Cambridge Toolkit, *Proceedings ESCA Eurospeech*, Rhodes, Greece (1997)
- [5] Papineni, K., Roukos, S., Ward, T. and Zhu, W. BLEU – a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics* (2002).
- [6] Weerasinghe, A.R. Bootstrapping the lexicon building process for machine translation between 'new' languages. *Proceedings of the Association of Machine Translation in the Americas Conference (AMTA)*, 2002.
- [7] Och, F.J., Tillmann, C. and Ney, H. Improved alignment models for statistical machine translation. In *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Maryland, 1999.



A Tool for the Management of ebXML Resources

S.S. Sooriarachchi, G.N. Wikramanayake, G.K.A. Dias
University of Colombo School of Computing

E-mail: sumekhala.sooriarachchi@ifs.lk

Abstract

The ebXML initiative is designed for electronic interoperability, allowing businesses to find each other, agree to become trading partners and conduct business. ebXML repository is used to store the ebXML resources and the ebXML registry is used to discover these resources. ebXML resources may be in the form of XML documents, Document Type Definitions, XML Schemas, UML models and various other forms. Therefore a proper tool is required for the management of these resources.

Management of ebXML resources has to be done by the experts who create and modify and are in charge of these resources. If there is a tool for the management of these resources, the experts can concentrate more on the content of these resources rather than management of them.

This paper describes the implementation of a graphical tool for the management of ebXML resources based on the proposed ebXML specifications. Two approaches are being used to discover these resources. Registry Navigator is one, which is in a tree structure and gives a full view of the registry. Query Manager is the other, which enables quick referencing to the resources if the user is fully aware of what resources are needed. The documents discovered in this manner are to be opened in the Editor pane of the tool. This editor facilitates the creation and modification of documents easily.

The tool also incorporates capabilities to add and remove resources through the Life Cycle Manager. It also has a mechanism, which keeps track of different versions of the resources, so that the preferred version can be referred with minimum effort. Finally, it allows non-expert users to dynamically access the content of the registry over the web.

Keywords: ebXML Resources, Registry, Repository, Graphical Tool

1. Introduction

Today the world of electronic collaboration [5] is developing rapidly, introducing new technologies, and new ways of collaborating. The success of collaboration will depend on the ability of a corporation to make sure that their applications are not only dynamic, but maintain a high degree of inter-operability with collaboration partners.

Electronic Data Interchange (EDI) [4] essentially defined the technology of electronic collaboration for the last millennium, but its popularity is waning for a variety of reasons. EDI is an expensive solution, due to the high cost of network infrastructure and system integration. It has also proven to be complex, difficult to maintain and inflexible in the face of changing market conditions. Smaller businesses with low volume collaboration needs, simply preferred to stay away from EDI.

The challenge to achieving dynamic e-business collaborations lies in the need to have a low cost, flexible software solution that allows corporations to build new applications in response to changing business needs while adhering to a defined electronic business standard.

Web services [9, 11] offer the potential for seamless application integration regardless of programming language or operating environment. Web services technology is based on a set of existing Internet standards and widely accepted specifications: HTTP, XML, SOAP, WSDL and UDDI. Web services alone are insufficient to achieve effective electronic collaboration unless Web services are applied in the context of collaboration standards such as ebXML [3].

ebXML is a set of specifications that enable a modular, yet complete electronic business framework. If the Internet is the information highway for electronic business, then ebXML can be thought of as providing the rules of the road. The ebXML initiative is designed for electronic interoperability, allowing businesses to find each other, agree to become trading partners and conduct business [7]. ebXML brings EDI's benefit of a common framework for conducting business in a supply chain

management model [15] to the small-to-medium-sized enterprise and to those enterprises that require a more flexible, loosely coupled e-commerce infrastructure.

1.1. Management of ebXML Resources

Even if the ebXML specifications exist, the benefits would not be fully realized if they are not properly managed or if they cannot be discovered as and when they are needed. Registry and repository is the mechanism to register and discover company and business service profiles, as well as business process specifications with related message exchanges, and other XML and e-commerce resources.

Early adoption of XML by industry partners is creating opportunity for information reuse and collaborations over the Web. At the same time, the rapid emergence of XML Document Type Definitions (DTDs) and vocabularies from industry and government sectors has focused attention upon issues of resource identification, classification, cataloging and delivery that hinder reuse and interoperability. The results of new collaborative endeavors are not necessarily easy to identify and access on the Internet.

For ebXML resources, registry/repository acts as a central warehouse. It is used to submit, store, retrieve and manage resources to facilitate ebXML-based business-to-business (B2B) partnerships and transactions. Submitted information may be, for example, in the form of business profile information, XML schema and documents, business process definitions.

Also there should be a discovery mechanism for businesses to find and engage one another. Registering a business service profile and business process schema in an ebXML registry/repository enables them to be located.

Since every company engaged in e-business concentrate on the growth of their individual enterprises, a central body is needed to initiate and be responsible for management of ebXML resources. Especially in Sri Lanka, where the businesses are still at the dawn of e-collaborations, it is better to have a responsible organization that can initialize and direct trading partners to use ebXML resources. This responsible central organization can build a comprehensive registry covering many areas of business giving easy access to these resources, which will facilitate the specifications to get a better recognition and a quick adoption. The work presented here is part of achieving such a goal.

2. Registry/Repository

2.1. Concept of Registry and Repository

Registries are aimed to manage the challenge of passing consistent information between a business system and its' suppliers and customers. Associated with this is the critical need to provide the means to accurately and quickly locate specific information on a topic or a domain.

Discovering new trading partners and the rules for engaging in a particular line of business are clearly powerful reasons for accessing e-business registries. But there are other functions designed to meet the needs of cost effective and timely use of information.

Essentially the functionality of an e-business registry can be divided into three broad domains.

- Providing a directory of members and services available with search and discovery.
- Providing human readable technical documentation and specifications organized using applicable domain classifications and categories.
- Enabling automated machine to machine e-business interactions through machine readable consistent content and process definitions, associations and linkages.

The information that can be discovered via the registries actually resides in the repositories. A repository is not just a passive data dictionary or database. It is an integrated holding area that should also keep the information up to date by providing processing methods and make it available to a user as needed. A repository, which maintains valuable information about all of the information system assets of an organization and the relationships between them, acts as a central manager of all of the information resources in an enterprise. A repository should provide services such as change notification, modification tracking, version management, configuration management, and user authorization [10].

The widespread availability of XML-capable clients and their flexibility in structuring information make it possible for XML to become the universal data format. Without the help of a repository, it will be difficult to control XML objects in a manageable way and make them available when needed.

XML repository provide several basic functions such as importing/exporting XML data from original text files, user check in/check out, version control, as well as searching and querying on repository items (XML documents). In the electronic commerce world, XML repositories are the online source for

obtaining the appropriate tag, document-type definition, data element, database schema, software code or routines. As a result, companies, especially small enterprises, can speed up processing and expand their ability to conduct electronic commerce [10].

2.2. ebXML Registry/Repository

ebXML registry/repository acts as a central warehouse for ebXML resources. It is used to submit, store, retrieve and manage resources to facilitate ebXML-based business-to-business partnerships and transactions. Submitted information may be, for example, in the form of business profile information, XML schema and documents, business process specifications, business context descriptions, Unified Modeling Language (UML) models, business collaboration information or even software components. Runtime artifacts of ebXML are shown in Figure 1.

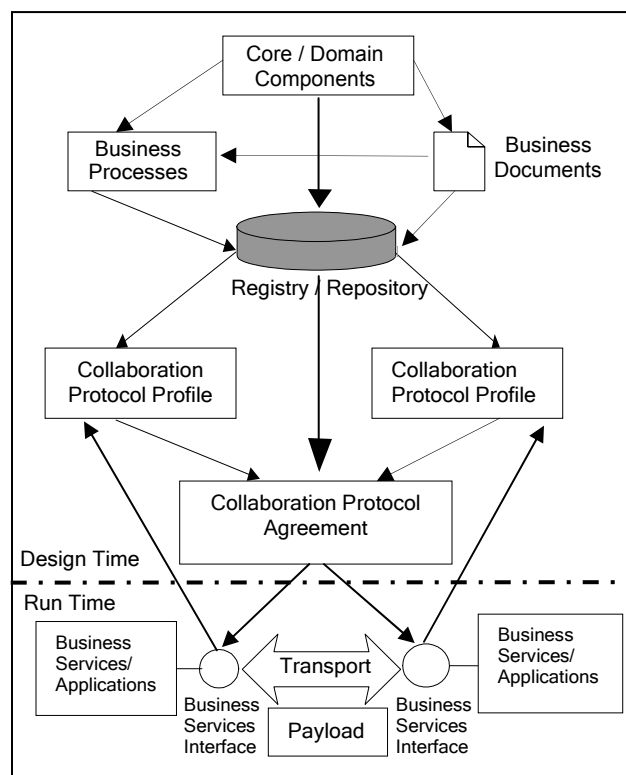


Figure 1: ebXML runtime artifacts [5]

A variety of XML formats are used in ebXML, such as:

- Business Process Specifications (BPS)
- Business Document Specifications (BDS)
- Collaboration Protocol Profiles (CPP)
- Collaboration Protocol Agreements (CPA)
- Log/Audit Trail interchange

- Core Components definitions
- UMM Models

Business processes represent the *verbs* of e-business. To enable integration of business processes within or between businesses, clear definitions of the business processes must be expressed in such a way, that they are understandable by the people and software of other businesses or business units.

ebXML defines a methodology for modeling business processes as a set of choreographed document exchanges, as well as the procedure for representing business process transactions in unambiguous ebXML business process schemas. These collaboration models make no demands on the underlying infrastructure. Consequently, businesses and industry organizations can develop and reuse business processes, without concern for the specific platform or software application that will execute the transactions.

Once a business process is defined, businesses need a standard means of describing the roles in which they are prepared to engage for that business process, as well as the technical capabilities they support to fulfill those roles. Generally, the description is defined in terms of roles such as *buyer* and *seller*. The CPP identifies which role or roles the party is capable of playing in each collaboration protocol referenced by the CPP.

CPP describes a partner's IT capabilities. These capabilities include what communication protocols (HTTP, SMTP, FTP etc.) they support, what security requirements they place upon the message exchanges, and what business processes they support. A CPP describes all the things a partner can do.

The protocol profile contains information about the business collaborations that a company supports and its message exchange capabilities. Using information from these profile documents, a collaboration agreement is formed to define the way in which parties will interact in the performance of business collaborations. A clearly defined trading agreement must be created that can be used to govern the transactions between partners.

A CPA first identifies the parties to the agreement. CPA contains following details:

- Communication protocols the parties will support.
- The messaging protocol to be used in exchanging business documents.
- Information needed to ensure a secure interchange of information between the parties.
- Business Transactions or services that the parties agree to interchange.

2.3. ebXML Specifications

The participants of ebXML had a vast amount of experience in various industries, EDI and XML standards and initiatives. They were able to bring their wealth of knowledge and experience to develop a set of specifications.

The ebXML framework consists of the following specifications [6]:

- ebXML Technical Architecture Specification
- Business Process Specification Schema (BPSS)
- Registry Information Model (RIM)
- Registry Services Specification (RSS)
- ebXML Requirements Specification
- CPP and CPA Specification (CPPA)
- Message Service Specification

"The RIM [12] provides a blueprint or high-level schema for the ebXML Registry. Its primary value is for implementers of ebXML Registries. It provides these implementers with information on the type of metadata that is stored in the Registry as well as the relationships among metadata Classes."

The RIM defines:

- Types of objects that are stored in the Registry
- How stored objects are organized in the Registry

A set of Registry Services that provide access to Registry content to clients is defined in the ebXML RSS [13]. The RSS defines the interface used to the ebXML registry as well as interaction protocols, message definitions and XML schema. The registry services permits access to the repository or content management system.

RSS assumes B2B exchanges that are carried out in the following sequence:

- BPS are submitted
- Business Process Documents are submitted
- Seller's CPP is submitted
- Buyer discovers the seller
- CPA is established after negotiations
- Once the seller accepts the CPA, the parties may begin to conduct B2B transactions

This specification also defines the actors who may interact with the registry, such as Registry administrator/Responsible organization, Registry user, Registry guest, Submitting organization (same as Registry administrator).

The ebXML Registry Service is comprised of a robust set of interfaces designed to fundamentally manage the objects and inquiries associated with the ebXML Registry. The two primary interfaces for the Registry Service consist of:

- A Life Cycle Management interface that provides a collection of methods for managing objects within the Registry.
- A Query Management Interface that controls the discovery and retrieval of information from the Registry

A registry client program utilizes the services of the registry by invoking methods on one of the above interfaces defined by the Registry Service.

2.4. Registry/Repository Systems

2.4.1. Sun ebXML Registry and Repository

The Sun ebXML Registry/Repository Implementation (RegRep) [17] can be used to submit, store, retrieve, and manage resources to facilitate ebXML-based B2B partnerships and transactions.

The RegRep implementation is based on open, non-proprietary, platform-neutral J2EE technology. What this means is that you can use the development tools, application servers, databases, and platforms you want. Core components of this implementation include a Registry Information Model, Registry Services, Security Model, Data Access API, Java Objects Binding Classes and JSP Tag Library.

2.4.2. OASIS XML Repository

The XML interoperability consortium OASIS has announced public access to the first phase of XML.org Registry, an open registry and repository for XML specifications and vocabularies [18].

The site is designed to both a central registry for XML schemas and other public resources (DTDs, namespaces, stylesheets, public key certificates), and an open development forum for designing useful repository/registry architectures.

The XML.ORG Registry was developed by Documentum and Sun Microsystems using software components from Documentum, iPlanet, and Oracle. Documentum 4i eBusiness edition, the content management platform powering the registry application, drives the entire process from the submission of a schema to its availability for public access via Documentum Site Delivery Services.

2.4.3. IBM XML Registry/Repository

The IBM XML Registry/Repository (XRR) [8] is "a data management system that manages and provides services for XML artifacts including schemes (DTD, XSD), stylesheets (XSL) and instance documents (WSDL). User can use XRR to obtain an XML artifact automatically, search or browse for an XML

artifact, deposit an XML artifact with or without related data, and register an XML artifact without deposit.

The registry provides a search of registered objects based on their metadata. Registry facilities include registration, search and retrieval of registered objects, and Administration.

The 'Repository' service "provides access to registered objects. Through the repository, a user can download a registered object using standard identifiers (URLs)." The current version of XRR runs on Windows NT, Windows 2000, Linux, AIX, and Solaris; it supports basic Servlet/JSP functionality. Databases: IBM DB2, Version 8, must be installed and running.

2.4.4. CENTRAL Registry Project

CENTRAL registry project of Boeing enterprise [2] provide a company-wide resource for registering, locating, sharing, and re-using XML schemas, DTDs, and other information needed to enable the electronic interchange of data and for understanding the meaning of that data.

2.4.5. RepoX XML Repository

RepoX, an XML repository [10], has been developed for the METEOR workflow system. It maps XML documents to a relational-object database and also provides extraction/retrieval, version control, check in/check out, and searching and query functions.

The RepoX repository provides full support for searching, querying, and versioning. An XML document can be modeled as a "rooted, directed, ordered, and labeled tree". To access and manipulate the XML document as a tree structure, the Document Object Model (DOM) core interfaces are used in the RepoX.

2.5. Graphical Tools and Web Interface

Since ebXML resources are stored in ebXML Registry Repositories there is a need for graphical user interfaces in order to be able to manage these resources. Our aim is to study the features required for an ebXML registry repository and then to design and implement a graphical tool and a web interface to manage these resources. For this purpose we have studied a number of graphical user interface techniques for general and specific management, editing, storage and validation of ebXML artifacts such as XML specification documents. Above listed registry repositories are part of this study.

3. Design of RRMS

The registry repository management system (RRMS) is designed in three main layers namely presentation layer, logic layer and the storage layer. The main architecture of the system is shown in the Figure 2.

The presentation layer contains the four interface components: *Life Cycle Manager* and the *Query Manager* with the navigator belonging to the standalone tool, and the web interface of the web based subsystem. Standalone tool is to be used only by the registry/repository admin of the responsible organization and the web-based system is for the trading parties from different industries.

The logic layer will take care of depositing the XML documents in its correct location in the repository, obtaining metadata from users about the documents to be saved and displayed in the registry. The storage layer physically represents the registry and repository. Design details of RRMS with class diagrams and sequence diagrams can be found in [16].

The repository is primarily categorized under different industries. Under each industry there would be a node for its' Business processes, CPPs of the trading parties of that industry and CPAs of the trading parties who get together for collaborations.

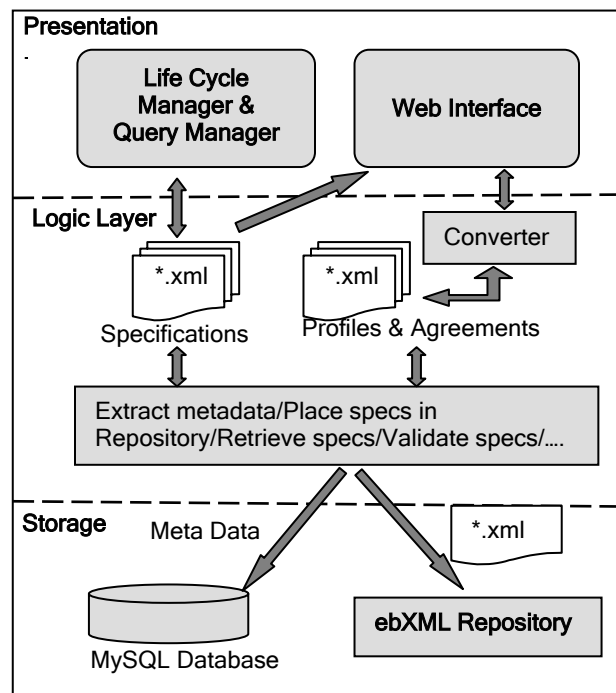


Figure 2: Main Architecture of RRMS

Generally there would be many business processes under one industry. Therefore the business

processes node has sub nodes for all the business processes corresponding to that industry.

A particular business process node further divides into BPS and BDS under this model. Since there would be more than one document that are exchanged between trading parties for a particular business process, there would be a node for each of the document specification.

In this repository structure, only the leaf nodes will carry repository items (i.e. specifications, profiles or agreements).

4. Implementation of RRMS

4.1. Environment

JBuilder 7 Enterprise Edition of Borland Software Corporation is used for the development of the graphical tool and the JSP pages of the web application. JBuilder contains major improvements in developer productivity, as well as a cleaner, more intuitive user interface and dramatic performance enhancements.

MySQL 4.0, the most popular Open Source SQL relational database management system, is developed, distributed and supported by MySQL AB. The MySQL Database Server is very fast, reliable, and easy to use. It also has a practical set of features developed in close cooperation with the users.

MySQL Connector/J 2.0.14 (Formerly MM.MySQL - Mark Matthews JDBC Driver for MySQL), which is a free product, is the JDBC driver used during the implementation.

4.2. Proof of Concept Implementation

This section gives a detailed description of the implementation phase of the registry/repository management system. As described under the design, only a selected number of resources are considered for the implementation. Implementation was carried out in the following steps.

4.2.1. Creation of the database: Registry

As the first step, the database, which plays the role of the registry in this system, was created. This database contains data about the documents stored in the repository, and also a reference (URL) to the absolute location of these resources.

The database named *registry* is created in the MySQL database server containing the following tables. The MySQL command used in obtaining the database structure is also given here. Tables in the

registry are *agreements*, *bds*, *bpps*, *businessprocess*, *cpemplates*, *cpptemplates*, *document*, *industry* and *profiles*.

Of the above, the tables *industry* and *businessprocess* and *bds* does not contain information about a specific document. They are needed for the relationships in the database and to maintain the repository structure.

4.2.2. Creation of folder structure: Repository

It is in the folder structure, the resources are actually kept. When creating each of the documents, it also dynamically creates the path (URL) in which the document is to be kept, according to the options selected by the user. For e.g. if the creation of a new version of a business document spec is considered, code segment is as follows. The full URL is sent to the database (*document* table) to be used when retrieving the documents.

4.2.3. Development of the Graphical Tool

Graphical tool plays an important part in the system. It is through this, the expert users manage the registry/repository. The main parts of the tool consist of registry navigator, life cycle manager, query manager, editor pane and web interface.

(a) Registry Navigator

The Registry Navigator was implemented using the Swing component, Jtree, which can be used to provide a view of hierarchical data. Like any non-trivial Swing component, the tree gets data by querying a data model. The tree in this case was placed in a scroll pane to allow easy navigation when the tree grows in size with the addition of more and more new resources to the registry/repository.

The tree is generated dynamically with the data from the appropriate tables in the database and by placing them in the tree so that it reflects the real folder structure of the repository. The tree is also refreshed, when a new resource is added, so it gives the most updated view of the registry all the time.

The documents, which are represented by the leaf nodes of the tree, can be opened in the Editor pane directly by selecting a document in the navigator and clicking on *Open* button (Figure 3).

(b) Life Cycle Manager

The Life Cycle Manager has to provide facility for the expert users to add resources to the registry/repository, modify these resources and finally remove these resources.

The addition of resources is done through the menu *Life Cycle Manager* of the graphical tool. How each of its menu items function is described below.

New Industry: Through this frame a new industry can be added to the database and the navigator when it is refreshed.

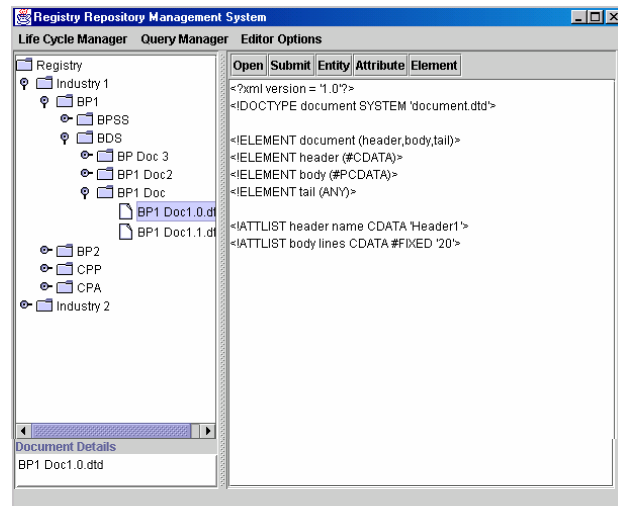


Figure 3: Registry navigator with a document opened

New Business Process: Through this a new business process is added to the database. The new business process is for a particular industry. Therefore, this frame enables selection of the required industry through a combo box. This combo box is filled with the industry names from the *industry* table.

New BPSS: Through this a new BPSS is added to the database. This is used to add a new business process specification schema for a selected business process of a selected industry. The available industries are added to the combo box by a query and the business processes are filled to another combo according to the selected industry.

New BDS: This has two sub menu items, one to add a New BDS type and the other to add a New Version. Here too, the new resource is added for a selected process of a selected industry. This selection is enabled through two combo boxes similarly to the above-mentioned methods. When a new BDS type is added *bds* table is updated and when a new version is added the 'document' table is updated.

CPP and CPA: Both the menu items function very much similarly. Only the tables that are queried and updated differ. For the CPPs *profiles* and

cpptemplates tables are used while *agreements* and *cpatemplates* are used for the CPAs. These are added to the registry, industry wise. When a new template is added it is sent to the *cpptemplates* table or to the *cpatemplates* table.

Removal of resources is enabled at a higher level, by giving the option to remove resources of a whole industry or an entire business process (Figure 4). When this is done, the related documents are also removed from the database according to the primary keys of the *industry* table and the *businessprocess* table (i.e. Delete is cascaded). Also removal of resources is enabled at an individual document level. That is, user can remove one document at a time. Before removing a resource, the system will display a dialog box to make sure that the removal is deliberate and not done by mistake.

In both these cases, the node has to be removed from the tree navigator and the corresponding records have to be removed from all the tables and finally the document has to be deleted from its exact location in the repository.

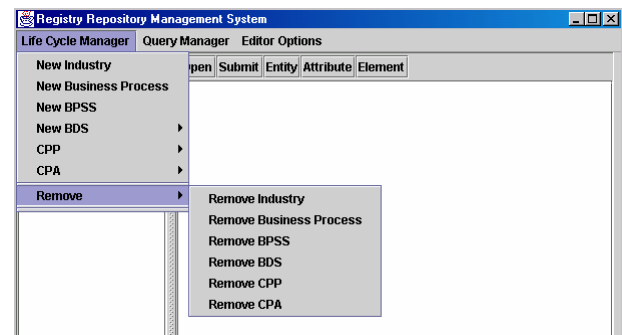


Figure 4: Removal of Resources

(c) Query Manager

Query Manager allows the expert users to query for BPSS, BDS, CPP and CPA.

Querying BPSS: The Query BPSS has two combo boxes for the user to select the industry and the business process. The business process combo box is filled according to the industry selected. The BPSSs that match the selected options are retrieved from the table *bpss* and displayed in a table. If the user wants to view a BPSS document, selecting a row in the table and then clicking on *View BPSS* button enable this.

Querying BDS: The functionality of this option is very much similar to the above explained one. Only difference is that the table queried is *document*, rather

than *bpss*. Figure 5 is the interface to retrieve BDS documents.

Querying CPP: Query CPP tabulates the details about the CPPs submitted to the registry. The CPPs can be retrieved industry wise through this query manager by selecting the required row from the table. The CPP template also can be retrieved according to the selected industry. In here, the tables queried are *profiles* and *cpptemplates* to retrieve the documents stored in the repository. Functionality is similar to that of Querying BPSS.

Querying CPA: This is very much similar to the Querying CPP, except for difference in the tables queried. They are *agreements* and *cpatemplates* for the case of CPAs.

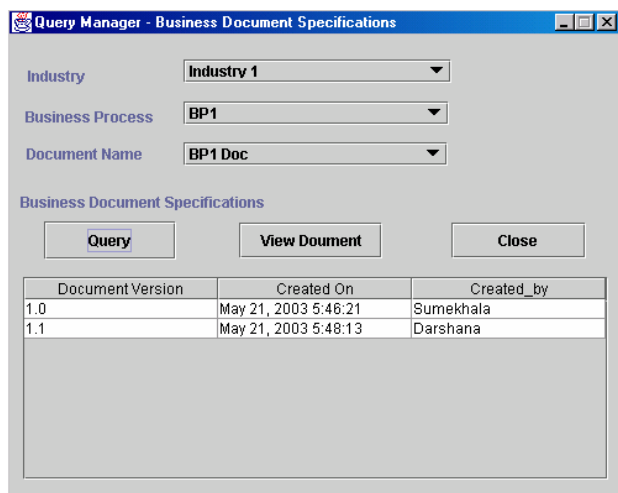


Figure 5: Accessing BDS using query manager

(d) Editor pane

Editor pane allows the expert users to add and modify the content of the resources created with the Life Cycle Manager menu options. First the documents have to be opened in the Editor pane. The URL of the currently opened file is kept in a String variable to be used when submitting the file back to the repository after modifying the content. To make it easy for the expert users to add content to the document files they are creating, the Editor Options menu gives some options to the users.

The options given are mainly to support creation of BDS documents. Since BDS documents are to be created in DTD format, the addition of building blocks of DTDs such as, a root, elements, attributes and entities are enabled through the sub-menu items.

(e) Web interface

The web interface is for the purpose of displaying the registry content and discovering BPS. Java Server Pages (JSP) was used to generate the dynamic web pages which provides easy navigation through the registry's resources.

Here also the *registry* database is queried according to industry and the business process selected by the user, by passing the primary key of *industry* table or *businessprocess* table as a parameter to the subsequent .jsp pages.

5. Evaluation

This section evaluates the achievements by comparing them with the predefined objectives for various stages of the project.

This project carried more work during the analysis and design stages. A comprehensive analysis had been carried out covering the issues related to e-business, e-business collaborations, standardizing e-business. Extra effort was put in understanding the ebXML framework and the role played by registry/repository. To understand the functions of registry/repository, a literature survey was carried out covering a number of related architectures. By putting together the information gathered as above, an in depth requirements analysis was undertaken achieving the objectives set.

During the design stage, architecture for the registry and repository had to be developed, along with the architecture for the tool as planned. According to the requirements and to reduce the complexity, the registry was designed as a database, which would store meta-data about the resources. The repository was designed as a folder structure as described in the design section.

The tool is designed to help the expert users, to manage the ebXML resources according to the defined scope of the project. The tool was designed to have a navigator for the purpose of discovering resources in the registry/repository. The resources are to be accessed and manipulated through the Life Cycle Manager and the Query Manager. These details are covered in the design section of this report. In this stage, in addition to the graphical tool required by the expert users, some other functionality, which are needed by the non-expert users such as submission of CPP and CPA are also designed having related future work in mind.

Proof of concept implementation for the above-mentioned design has four main parts. The Navigator in the form of a tree, which facilitates discovery of resources, the Life Cycle Manager to support the management of the resources through their life time,

the Query Manager for discovering and retrieving the resources and the Editor pane, which supports the creation, modification and viewing of the resources. Also a web application is implemented as expected, to discover the registry content, which are updated dynamically over the web.

The implementation of submission of CPPs and CPAs is not covered since it is out of scope of the targeted work. The design done in this area is expected to direct future work related to providing registry facilities to the non-expert users.

6. Conclusion

The Graphical Tool for the management of ebXML resources designed under this project was targeted towards the expert users who would be using the registry/repository, for the purpose of maintaining these resources.

The suitability of such a tool for the e-business community of Sri Lanka was a main concern. Since e-business is still at a primary stage in Sri Lanka, a centralized tool was thought to be more suitable, where initially one authority would maintain the ebXML resources through the resulting tool.

During the implementation of the tool, only a limited number of ebXML resources were taken into consideration, namely business process specification schemas, business document specifications, collaboration protocol profiles and agreements. And also limited Editor options were given to avoid 'reinventing the wheel'.

By developing an architecture for the registry/repository, and designing a graphical tool to manage the ebXML resources to be stored in the registry/repository with a supporting proof of concept implementation the objectives of the project have been achieved within the given constraints. This tool is expected to initiate standardization of e-business in Sri Lanka and to cater to the needs of its' fast growing e-community.

6.1. Future work

The Graphical tool for the management of ebXML resources developed under this project can be thought of as a starting point for a number of important areas related to the development of e-business in Sri Lanka.

The graphical tool developed under the project has further capacity for enhancements. Following are some suggestions for improvements of the tool and to make it more effective and comprehensive.

- The Editor pane of the tool, which is to be used by the expert users to create the documents, can be improved by providing a wide range of Editing options as in well-known XML editors

like 'XML Spy' [1]. Wizards for creating such documents would be a value adding option to the tool.

- The registry can be made more comprehensive to include resources of various types, other than the resources considered for the purpose of this project. These other resources can be UML models, Core Components, etc.
- For the registry to serve the business community of Sri Lanka, it should contain well-prepared standards (ebXML resources) for various industries. This has to be done after a well planned careful analysis of each industry and then creating and submitting these resources to the registry/repository through the tool. A dedicated team is proposed to do such analysis and for the management of the registry/repository, so that it would increase the confidence the business personnel has on the standards.

Acknowledgements

This work was done at University of Colombo School of Computing are part of collaborative research with the OpenXML Laboratory, University of Stockholm. Valuable advice and guidance given by Mr. Anders W. Tell, Mr. Erik Perjons, Mr. Harsha Wijewardena and the members of the research group is acknowledged.

References

- [1] ALTOVA - XML Development, Data Mapping, and Content Authoring, accessed on 071103, <http://www.xmlspy.com/>
- [2] Breininger Kathryn, "CENTRAL Registry Project", Boeing aviation integration, e-business solutions, Standard Services Group, 2001, accessed on 071103, <http://boeingicp.eep.gmu.edu/presentations/Kathryn%20Breininger%20-%20Central%20Registry%20Project.pdf>
- [3] ebXML - Enabling A Global Electronic Market, accessed on 071103, <http://www.ebxml.org>
- [4] EDI Standards, Federal Information Processing Standards Publication 161-2, 1996, accessed on 071103, <http://www.itl.nist.gov/fipspubs/fip161-2.htm>
- [5] Electronic Collaboration: A practical guide for educators, Brown University, 1999, accessed on 071103,

- <http://www.lab.brown.edu/public/pubs/collab/elec-collab.pdf>
- [6] Harvey Betty, "The Role of XML in E-Business", Electronic Commerce Connection Inc., 2002, accessed on 071103, <http://www2.cs.uregina.ca/~tang112x/research/papers/2003w/>
 - [7] Ibbotson John, ebXML Trading-Partners Specification, Internationales Congress Centrum, XML Europe, Germany, 2001, accessed on 071103, <http://www.gca.org/papers/xml europe2001/papers/html/s09-2.html>
 - [8] IBM alphaWorks Releases XML Registry/Repository Data Management System, 2001, accessed on 071103, <http://xml.coverpages.org/ni2001-06-04-a.html>
 - [9] INCITS — the InterNational Committee for Information Technology Standards (formerly X3), accessed on 071103, <http://www.X3.org>
 - [10] Minrong Song, John A. Miller and Ismailcem B. Arpinar, "RepoX: An XML Repository for Workflow Designs and Specifications", Technical Report #UGA-CS-LSDIS-TR-01-012, University of Georgia (August 2001) 43 pages, accessed on 071103, http://chief.cs.uga.edu/~jam/home/theses/song_thesis/song_minrong_repoX.pdf
 - [11] Morais Pravin, "Dynamic e-Business Using Web Service Workflow", SearchWebServices, New, June 2002, accessed on 150903, <http://www.cysive.com/news/062602.htm>
 - [12] OASIS/ebXML Registry Information Model v2.1, Approved Committee Specification - OASIS/ebXML Registry TC, 2002.
 - [13] OASIS/ebXML Registry Services Specification v2.1, Approved Committee Specification - OASIS/ebXML Registry TC, 2002, accessed on 071103, <http://www.oasis-open.org/>
 - [14] Open ebXML Laboratory- project catalog, accessed on 071103, <http://www.openebxml.org/>
 - [15] Smith William C and Etelson David J, "e-Business XML for Global Purchasing and the Supply Chain", International Federation of Purchasing and Materials Management (IFPMM) World Congress, South Africa 2001.
 - [16] Sooriarachchi S.S., "A Graphical Tool for Management of ebXML Resources in Registry/Repository", B.Sc. Dissertation, University of Colombo School of Computing, May 2003.
 - [17] Sun ebXML Registry and Repository Implementation, accessed on 071103, <http://www.ohelp.com/samples/xml/repreg/repreg-intro.html>
 - [18] XML.ORG Goes Live with First Phase of Open Registry & Repository for XML Specifications, 2000, accessed on 071103, <http://lists.oasis-open.org/archives/announce/200006/msg00011.html>



Asymmetry in Facial Expressions: 3D Analysis using Laser Rangefinder Systems

Pujitha Gunaratne^{†‡}, Nihal Kodikara[†], Yukio Sato[‡]

[†]Department of Communication and Media Technologies
University of Colombo School of Computing
35, Reid Avenue, Colombo 7, Sri Lanka.

[‡]Department of Electrical and Computer Engineering
Nagoya Institute of Technology
Gokiso, Showa, Nagoya 465-8555, Japan

E-mail: pug@ucsc.cmb.ac.lk, ndk@ucsc.cmb.ac.lk, sato@nitech.ac.jp

Abstract

This paper presents an effective approach to analyze asymmetries in facial expressions in three-dimension using range data. Five basic facial actions are captured on ten different subjects and their shape information is recorded with corresponding texture image. To analyze the range surfaces, which have been deformed characteristically during each facial action, pre-designed symmetric generic face mesh is adapted to measured range data using the least squares approximation method. Then mirror image triangular patches in left and right sides of the altered symmetric mesh are analyzed for each facial action to determine the degree of deformation. A simple variation computation method for the triangular patches in 3D is presented for the analysis of deformation. To illustrate the robustness of estimation, we compare the results of normal subjects with a patient with facial nerve paralysis disorders. High accuracy range data were obtained for the robustness of analysis using two high-speed rangefinder systems and their distinct configurations are discussed with potential impact on asymmetry analysis.

1. Introduction

The interests in generating computer assisted human facial models have been seen in the past, aiming at representing the human face and its components primarily for animation tasks [1] [2]. Most of such works extract the human faces from intensity images and construct 3D models by representing them in crude polygons. With the advent of acquisition methods of accurate 3D data by rangefinder systems and the development of sound modeling techniques, the application areas of human facial modeling have

modeling have developed rapidly. Unlike in early animation tasks where conventional 2D characters were a commonplace, the ability to generate accurate 3D facial models has flourished new demands in diverse application areas. Such areas can be categorized from cinematic animation, teleconferencing, and virtual reality, where realistic virtual avatars are of demand, to robust recognition and interpretation in analytical applications. The analytical applications process the acquired data to extract certain features form raw data for secondary interpretations. Some of the primary analytical applications belong to security and medical fields. Thus, in this work we present the usage of 3D facial data in an analytical application that determines the degree of asymmetry in facial expressions. The outcome is directly applied in a medical application to determine the degree of paralysis observed in expressions in patients with facial nerve paralysis disorders. Other potential application areas are recognition and identification in security and authentication environments.

Range sensing systems, which are capable of simultaneously acquiring 3D as well as color texture data, have been designed and presented in recent years [3][4][5]. The sensing units of the system in [3] rotate around the measuring profile in 360° while projecting laser stripes. It records data points in a cylindrical coordinate system and then derives X, Y, Z, coordinate locations using a secondary mapping algorithm. Suenaga et al. [5] presented a system based on triangulation, which projects slit-rays on to the profile. Both systems take about 15 second measuring time to acquire a complete upper part of the body.

The measurement time makes a significant impact on the robustness of the results, since the subjects in these systems are humans. They do not intend to make any movements during the measuring process and sup-

posed to be still until the entire process is over. Hence, longer the time it takes to measure, more the burden on the profile, and lesser the stability gained in measured data. The studies have revealed that as little measurement delay as 10 seconds is enough to put burdens on humans to make subtle movements during holding up the expressions, causing vulnerability in interpretations.

Therefore, taking these facts into consideration, here we introduce the application of non-contact high-speed semiconductor laser rangefinder systems that have successfully eliminated the problems identified above, for the accurate measurement of facial expressions. One system is the Minolta VIVID 910[®] non-contact 3D scanner, which is newly released to the market and installed at the Visual Computing Research lab of the University of Colombo School of Computing, and the other is called the CUBICFACER[®], which is developed at Nagoya Institute of Technology, Japan. Both systems equipped with laser scanners and color CCD cameras to acquire high-density range and color texture images. They employ triangulation method [6] for faster measurement. The estimated measuring time of these systems lies well within one second and thus they are ideal for facial expression analysis tasks.

The analysis of asymmetry in expressions is based on adapting a symmetric generic mesh with face topology on to the measured 3D shape of each facial action. Various approaches to adapt generic face meshes to deformed 3D surfaces are proposed in the past that applied in parameterized and control point models [7] [8] [9], spline based models [10] etc. In most cases, mesh adaptation require segmentation of underlying 3D surface or setting up control points on feature boundaries, thus generating overheads in processing. Our adaptation process of the generic face mesh consist of extracting pre-determined mapping points from the intensity images of each action that has taken at the same time as 3D shape data. With the adaptation of mesh to different facial expressions, we obtain deformations in the mesh elements. The degree of deformation in identical patches on left and right sides are thus compared to determine the asymmetries.

This paper consists of two major parts. The first part consists of sections 2 and 3, which introduces the design and implementation strategy of the rangefinder systems and acquisition of 3D facial data. The second part explains the proposed method of asymmetry computation for captured facial expressions. Section 4 describes the computation method, followed by the section 5 that analyzes the results of subjects measured.

2. Face data measurement

As discussed in the previous section, a face measurement system must have the capabilities of acquiring 3D shape and color texture data in high accuracy and speed,

in order to satisfy the correctness of estimations. The rangefinder systems we introduce here are based on the consideration of following design issues.

- (a) Capable of obtaining both 3D shape and color texture data.
- (b) Acquisition of occlusion free frontal face images.
- (c) Dense 3D data with minimal average error. (high accuracy).
- (d) High-speed in measurement (since human expressions are involved).

In the issue (a), the goal is to capture 3D depth information, as well as color texture data for each measured point on the object surface. This has been an important issue in most graphic applications, where the correct texture and shape information of each measured point is necessary for realistic modeling purposes. The issue (b) is concerned with capturing an occlusion free face data from frontal direction. When we measure a face from one view direction, some data holes may exist especially in the underside of the chin and in the nostril areas due to non-projection of laser beams. This problem is addressed with acquisition of multiple range images from different view directions, so that occlusions in one direction will be compensated by the images captured in other direction. The issue (c) addresses the accuracy of measured data on which preceding estimation computations are based. For robust analysis of asymmetry in expressions, accurate and dense sets of facial data are necessary, since some facial movements are very subtle. The issue (d) is the most important and the most resolute, which addresses the requirement of the speed. As mentioned in the previous section, the design criteria of the systems are mainly based on capturing a profile as faster as possible as to put least constraints on the subject. Implementing sound techniques to achieve the high-speed in measurement compensates the constraints in data acquisition on subjects.

In order to meet these requirements, we have selected two range scanning systems that are installed at our research facilities. The first system, which is depicted in Fig. 1 is a Minolta VIVID 910[®] 3D digitizer that consists of a CCD camera with three replaceable lenses and a laser scanner that projects a laser beam during scanning. Three lenses are for wide-angle, middle-angle and telescopic distance measurements. The replaceable cap is mounted in the upper part and the laser scanner is installed in the lower part of a vertically mounted arrangement as in Fig. 1. A rotating stage is also coupled to the system to measure medium sized objects. Thus different viewpoints can be obtained by rotating the stage by arbitrary angles, which can be totally controlled by software.

The second system, which is depicted in Fig. 2, is called the CUBICFACER that consists of an optical unit

and a laser scanner unit. The optical unit consists of a color CCD camera (CN-411, ELMO) and a diffuse light source, while the laser scanning unit consists of two laser scanners. The laser scanners are mounted on “V” shaped holding bars and the CCD camera is placed right in the center of the scanners, where two holding bars are joined. A white fluorescent light source is placed just above the camera and projected to a white wall in the background in order to generate diffuse lighting for imaging system. Use of diffuse light is necessary to avoid the formation of highlight on the measuring surface.



Fig 1. Minolta VIVID 910[®] 3D Imaging System

The color CCD cameras mounted on both systems play two roles. One is the capture of color texture image and the other is the involvement in shape measurement. Hence both systems are capable of recording texture as well as shape information of the measuring surface, satisfying the design issue (a). The Minolta VIVID 910[®] system consists of a rotating stage, which can be controlled by image acquisition software, to change the viewpoint of the objects to measure occluded areas in the previous measurements. For the measurement of facial expressions, this set up is altered slightly by placing a rotating chair in the place of the stage. Since we are interested in occlusion free frontal face images, three range images taken at front, left and right sides are sufficient for occlusion removal.

The CUBICFACER[®] system consists of two laser scanners. Each scanner projects laser patterns on to the face from either side of the camera complementarily. Two scanners are placed in such a way that each of them is able to measure one half of the face completely. Hence, after a single complete measurement, one complementary pass by each scanner, we get two range images correspond to the left and the right halves of the face. A simple integration of these two images provides a complete occlusion free-range map of the entire face image. Thus both systems satisfy the design issue (b).

The accuracy of measurement in the Minolta VIVID 910[®] system is measured as 0.22mm in X direction, 0.16mm in Y direction and 0.10mm in Z direction with the telescopic lens in fine mode, which is the standard

mode of Minolta specifications. In similar measurements, CUBICFACER[®] system produces minimum depth error of 0.5mm with objects placed at 300mm measuring distance. According to these statistics, both systems satisfy the requirement (c) with acceptable accuracy for the measured data.

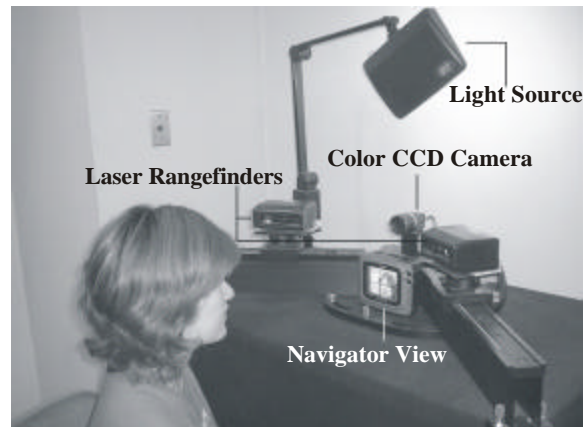


Fig 2. CUBICFACER[®] Face measurement system.

The laser scanner of the Minolta VIVID 910[®] system uses class 2 (60825-1), with “Eye Safe”, Class 1 (FDA) lasers. The measurement time is documented as 0.3 sec. in the fast mode, 2.5 sec in the fine mode and 0.5 sec. in the color mode. Two laser scanners of the CUBICFACER[®] system employed with semiconductor lasers of 680 nm, 40 mw power dissipation, a cylindrical lens and a polygonal mirror with 12-faces, which rotate at 3600 rpm. Laser slit rays are beamed to the polygonal mirror through a cylindrical lens with switching the laser at proper timing intervals according to encoded stripe patterns. Thus, spatial light patterns that are generated by temporal switching patterns strike on the surfaces of the measuring object. Thus a range image can be obtained in 17/60 seconds with each scanner, and a color image in 1/30 seconds. Therefore, both range and color images of a human face can be measured in 36/60 seconds (0.6 sec.) at the highest speed. Thus, speed produced by both systems well within the satisfying limits of the design issue (d). In both systems, since a single camera is used at a constant view direction with respect to the scanners, captured intensity and range images correspond to each other by pixel order. Hence the integration process of two images is faster than other range-finder systems. Since two systems do not differ much compared to their specifications, in the preceding sections we describe the measuring and analyzing techniques of asymmetry measurements in facial expressions for data acquired by the CUBICFACER[®] system.

2.1 Measurement technique

Since the measuring time depends on the measuring technique used, the CUBICFACER[®] system employs a faster space-encoding measuring technique to capture 3D data. The space-encoding method is proved to be one of the fastest techniques available, when using an ordinary video camera as an image recording device. More details on this principle can be found in the materials [11] and [12].

2.2 Measurement Procedure

The measurement procedure of this system can be categorized into following steps. Note that the numerical values embedded in parenthesized correspond to the actual processing time of each step for on an average person.

- Step 1: Turn the light source on before the measurement.
- Step 2: Capture the color image and then turn the light off (1/30 sec.).
- Step 3: Capture the range images with left and right laser scanners simultaneously (17/60 sec. per scanner).

In the first step white light source is turned on before the measurement starts. This step is to prevent the subject from moving. Because in a dark room environment, a person might move by a reflex, if laser patterns are projected onto him suddenly. Another reason is to adjust the correct position of the face for the imaging system. Moreover, by doing so beforehand would shorten the measurement time due to the reason that the light source takes time to turn on. In the next step, a color image is obtained and the light source is turned off. Then both range scanners operate simultaneously to capture the range images.

The total measurement time is the combination of processing time and the lag time between each step. If there is no lag time in switching steps, both range and texture images could be obtained in 36/60 seconds (0.6 sec.). However in practical situations, the measurement time is recorded about 1.0 seconds.

3. Construction of 3D Face model

The 3D image of the frontal face is generated by integrating left and right partial range images (Fig 3b) on their common boundary. In general, the process of range image integration is a tedious task, which involves extensive computations. But in this system, since a single camera is used to measure both range images at a same view direction, the integration task is simplified to taking the average range values along the common boundary of the overlapping regions. A weighted average method is used in the averaging process, which assigns

larger weights to the most dense data areas between two images [11] [12] for the combined image.

Despite the use of weighted averaging method, some unmeasured areas still exist in the image, mainly due to the black areas of the face. These include the eye-brows, pupils of the eye, etc., since they do not reflect back the projected laser patterns sufficient enough for the camera to record a measurement. Thus these areas are filled with a linear data interpolation method based on the neighborhood pixel values. Finally we could obtain a smooth range image of the face with 512 x 242 resolution (Fig 3c).

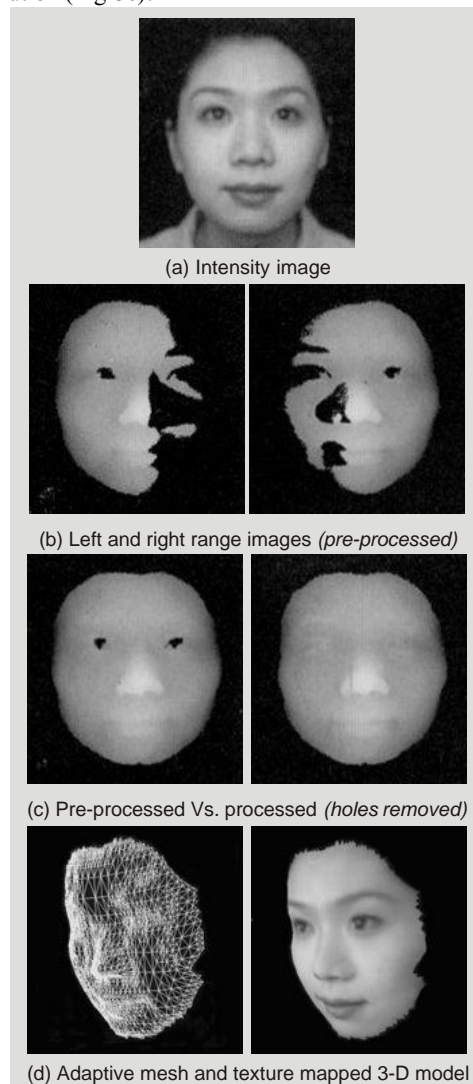


Fig 3. Formation of 3D frontal face model

An adaptive mesh algorithm is applied to the processed 3D face range data and then color texture is mapped to generate a realistic 3D face model as displayed in Fig 3d. This model can be rotated along X, Y,

and Z axis to change the viewpoint as shown in Fig 4. It can be seen that the constructed model retains the 3D structure and the color texture information at any view-able direction in 3D space.

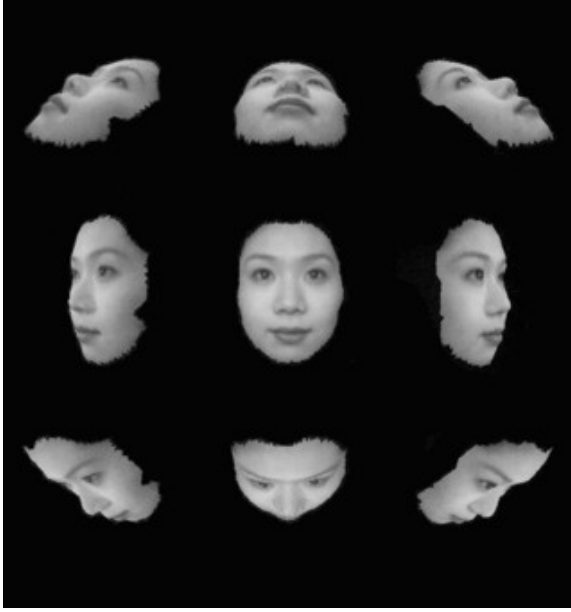


Fig 4. Face model in 3D view space

Hence it is simple to generate 3D models even when the view point is shifted or the light source direction is changed.

4. Asymmetry computation

The adaptive mesh generated in the texture mapping process does not possess consistent triangle density on both sides of the face (Fig.3d). Therefore it is hardly suitable for interpretations based on symmetry features of the face. Thus, we adopt apre-designed arbitrary generic mesh (Fig. 5) that is symmetric along the median plane, which is the vertical plane passes through the center of the nose, and cuts the face into identical left and right halves.



Fig 5. Generic face mesh

This generic mesh is in the 2D from, lies on the XY plane. Therefore we adopt a method of wrapping the mesh on to the measured 3D range data with the use of the information in the corresponding color texture image.

4.1 Mesh adaptation

The mesh adaptation can be time consuming, tedious process if it involves segmentation of range data to extract facial features. Instead, here we apply a simple method of extracting features by using the corresponding color image, since it possesses the property of one-to-one correspondence with the range image. We select 42 pre-determined points on the color image manually, which correspond to mapping points on the face mesh (Fig. 6), for the adaptation process.

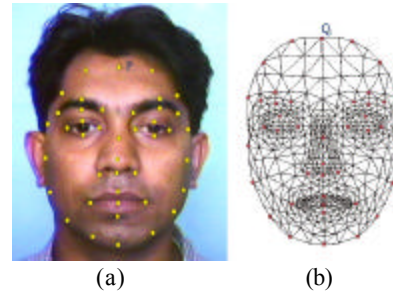


Fig 6. Points selected for mesh adaptation. (a) Texture points (b) Identical mesh nodes.

Points extracted from the color image are then mapped to the corresponding mesh nodes by approximating a polynomial function using a least squares estimator.

4.2 Least squares approximation

An N^{th} order polynomial function is approximated to the range data using a least squares estimator to adapt the generic mesh to the measured data. This process consists of two steps. First we move the mesh vertices to the extracted points of the color image, and then map the Z values from the corresponding range image.

Let us consider the parametric function given by,

$$Z = f(x, y).$$

Where, $f(x, y)$ represents by a polynomial of N^{th} degree, given by,

$$f(x, y) = a_{00} + \sum_{j=1}^N \sum_{i=0}^j a_{j-i,i} x^i y^{j-i} \quad (1)$$

When $N=2$, it takes the form,

$$f(x, y) = a_{00} + \sum_{i=0}^1 a_{1-i,i} x^i y^{1-i} + \sum_{i=0}^2 a_{2-i,i} x^i y^{2-i} \quad (2)$$

Now consider match points P_i and Q_i where $i=1, \dots, n$, represent points on the color image and the mesh respectively (Fig. 6). P_i 's are extracted from the color image and Q_i 's are known with respect to the topological mesh.

Let (x_{P_i}, y_{P_i}) and (x_{Q_i}, y_{Q_i}) represent 2D coordinates of P_i and Q_i respectively. We can thus calculate the displacement vectors, $\mathbf{dx}_i = (x_{P_i} - x_{Q_i})\bar{\mathbf{m}}$ and $\mathbf{dy}_i = (y_{P_i} - y_{Q_i})\bar{\mathbf{n}}$, where $\bar{\mathbf{m}}$ and $\bar{\mathbf{n}}$ are unit vectors along x and y direction respectively, for all matching points $i=1, \dots, n$. Since we extract 42 points for initial matching, n is set to 42. To calculate displacement vectors for the entire data set of the mesh, we approximate the parametric function given in eq.(2) using the least squares method, polling \mathbf{dx}_i and \mathbf{dy}_i in Z axis as depicted in Fig. 7.

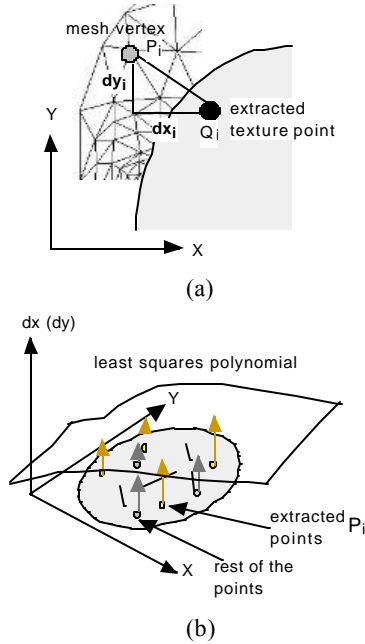


Fig 7. Least squares polynomial approximation.
(a) Displacement vector calculation. (b) Polynomial approximation.

Thus, coefficients $a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02}$ can be calculated using displacement vectors and therefore the nodal displacements of other points can be calculated by simply interpolating the polynomial function.

We repeat this procedure again by increasing the order of the polynomial to move the mesh points further closer to the expected locations by iterative approximation. We then separate feature points on different regions of the face, namely eye, nose and mouth regions, where a high concentration of facial features is observed.

This local matching is done to ensure a better mapping for the prominent feature areas of the face. Finally, Z values are mapped from the corresponding depth values of range data. Since both texture and range images have one-to-one correspondence, direct mapping is possible without any transformations. Once the 3D mesh is generated, we apply asymmetry measurements against the measured facial actions to estimate the difference of deformation on both sides of the face.

4.3 Estimation of Facial deformation

Facial deformation is estimated by calculating the variations in mirror image patches of left and right sides of the adapted generic mesh for each measured facial action.

Patch variations are estimated with respect to the sub-meshes representing different regions of the face. Forehead, eye, nose and mouth meshes are defined in the generic face mesh beforehand, and used to estimate the variations. Consider two matching mirror patch pairs in a given sub-mesh marked as P_{Li} and P_{Ri} , representing left and right side patches respectively (Fig. 8). Their corresponding edge lengths are denoted as ξ_{Li} and ξ_{Ri} respectively, where $i = 1, 2, 3$.

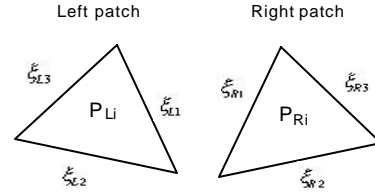


Fig 8. Matching mirror patch pairs.

If the variation of the i^{th} patch is \mathbf{S}_i^2 and total path variation of a given sub-mesh with N patches is \mathbf{S}^2 , we can define,

$$\mathbf{S}^2 = \sum_{i=1}^N \mathbf{S}_i^2 = \sum_{i=1}^N \sum_{j=1}^3 \left\| \mathbf{x}_{Li_j} - \mathbf{x}_{Ri_j} \right\|^2 \quad (3)$$

Thus, we measure the variations in terms of change in edge lengths of sub-mesh patches. In the applications of asymmetric facial expression analysis, it is often required to measure the relative variations of different facial actions and compare between them (Fig. 9).

In a similar calculation, as done in the previous case, suppose patch P_{Li} of expression A00 occupies the patch P'_{Li} in expression A01. Let the patch variances of left and right sides denote \mathbf{S}_L^2 and \mathbf{S}_R^2 respectively.

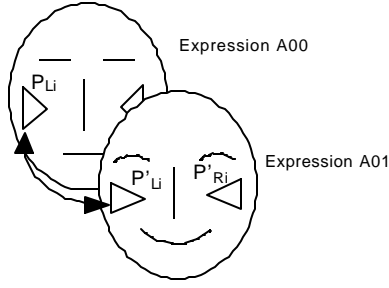


Fig 9. Patch Variation comparison of two expressions.

Thus,

$$s_L^2 = \sum_{i=1}^{N_L} \sum_{j=1}^3 \| \mathbf{x}_{Li_j} - \mathbf{x}_{L'_{ij}} \|^2 \quad \text{and}$$

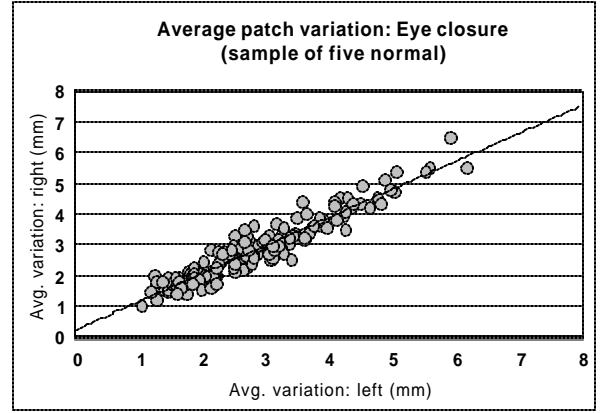
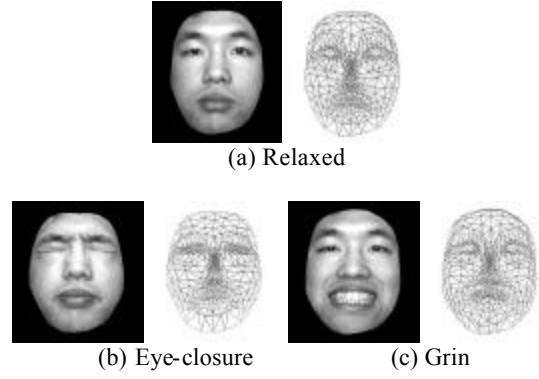
$$s_R^2 = \sum_{i=1}^{N_R} \sum_{j=1}^3 \| \mathbf{x}_{Ri_j} - \mathbf{x}_{R'_{ij}} \|^2 \quad \text{where } \mathbf{x}_L, \mathbf{x}_{L'}, \mathbf{x}_R, \mathbf{x}_{R'}$$

represent the lengths of the same patch in left and right sides in different expressions. N_L and N_R represent the number of patches in left and right sides of the same sub-mesh. Then the comparison is performed for both sides of the face to detect the asymmetry.

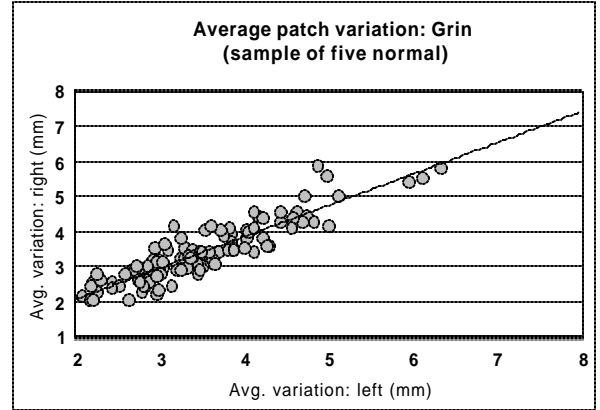
5. Results and Analysis

In the experiment we measured two facial actions, namely Eye-closure and Grin, of five different human subjects with no apparent expression disorders. We also measured the same expressions in patients with facial nerve paralysis disorders. These two actions Eye-closure and Grin are chosen for the results demonstration since they cover most of the movements of the face. We analyzed the range data of each action of all subjects measured by the CUBICFACER[®] rangefinder system, using above described variation estimations.

The subjects are first measured at the relaxed expression and then asked to generate each action and hold it up for about two seconds, during which the entire measurement procedure is completed. Patch variations are calculated for the Eye-closure and Grin actions with respect to the relaxed condition, thus enabling it to compare relative movements in facial parts in respective deformations. Distributions of normal subjects are averaged to form a control group and patient variations are compared against the control group during analysis. The mesh deformations of Eye-closure and Grin expressions of a normal subject and the correlation between left and right sides of the control group (average of five normal subjects) are given in Fig. 10.



(d) Correlation of Eye-closure of the control group



(e) Correlation of Grin of the control group

Fig 10. Correlation of the control group

Similar results of a patient for the two expressions are depicted in Fig. 11.

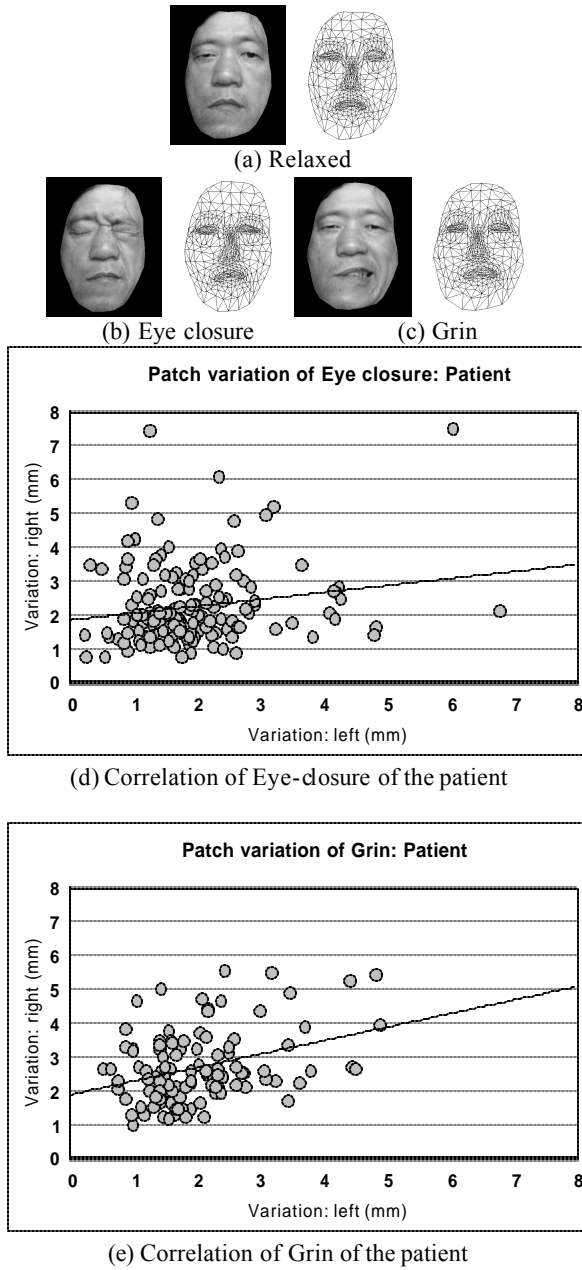


Fig 11. Correlation of a patient.

The standard error of the correlation distributions between the control group and the patient is compared and depicted in Table 1 below.

Table 1: Standard error of the distributions

	Standard Error	
	Eye closure	Grin
Control group	0.305	0.356
Patient	1.850	1.264

6. Conclusion

In this work we have demonstrated a method to analyze the asymmetries in facial expressions with the use of CUBICFACER[®] rangefinder system. The measurements are done in 3D with laser scanning. The resultant 3D range maps are high precision and able to acquire in about a second. This efficiency greatly contributes to the stability of measured data, thus eliminating the drawbacks that can be observed in other similar systems, as discussed previously. Although we have done the experiment in the CUBICFACER[®] system, similar results can be proven with the Minolta VIVID 910[®] range scanner, since both have similar high precision characteristics.

Some semiconductor lasers used in optical systems make harmful effects to the human body especially the retina when the rays hit directly at sensitive regions. There have been extensive testing that are carried out in the design phase of the CUBICFACER[®] system and the laser products used here passed the Japanese Radiation Safety Standards for laser products, which are regulated in JIS C 6802. There the safety standard is regulated with the MPE (Maximum Permission Exposure) parameters (J/m^2). The MPE parameters are classified by the wavelength of laser, exposure time, radiation form, observation state etc. The test results of such parameters of this system can be found in the document [13]. Similar testing for the laser safety has been done with the Minolta VIVID 910[®] system and it has passed the class 2 (60825-1) tests with "Eye Safe" Class 1 (FDA) lasers validations. Therefore both systems can be used with guaranteed safety.

The proposed method of calculating asymmetry was done by estimating the variations of identical patched in different expressions with respect to the initial rest position. Since both systems generate high density stable range data, satisfactory level of details in asymmetries are obtained for different subjects. Thus these systems can effectively be used with the proposed method of asymmetry estimation in demanding applications such as degree of facial paralysis detection, face identification and recognition applications where high precision is expected.

7. Acknowledgements

We would like to thank Japan International Corporation Agency (JICA) for providing grants for the Minolta VIVID 910[®] 3D scanning system for the Visual Computing Research group of the University of Colombo School of Computing. We also thank Dr. Seiichi Nakata of Nagoya University Hospital of providing patients for the much needed data acquisition phase and his keen interests in developing a system for clinical use.

8. Reference

1. F.I. Parke, "Computer Generated Animation of Faces. Proceedings, ACM annual conference", August 1972, pp. 451-457.
2. S.M. Platt, N.I. Badler, "Animating Facial Expressions", Computer Graphics, Vol.15, No.5, 1981, pp. 245-252.
3. Cyberware Laboratory Inc., "4020/RGB 3D Scanner with Color Digitizer", Monterey, CA, 1990.
4. M. Nahas, H. Huitric, M. Rioux, J. Domey, "Facial Image Synthesis using Skin Texture Recording", Visual Computer(6), Springer-Verlag, 1990, pp. 337-343.
5. Y. Suenaga, Y. Watanabe, "A Method for Synchronized Acquisition of Cylindrical Range and Color Data", Trans. Of IEICE, 25(12), 1991, pp. 3407-3416.
6. P.J. Besl, Active optical range imaging sensors, in Advances in Machine Vision, Ed. J. L. C. Sanz, Springer-Verlag, New York, 1988.
7. F. Parke. "Parameterized model for facial animation. IEEE Computer Graphics and Applications", vol. 2(9), 1992, pp. 61-68.
8. M. Oka, K. Tsutsui, A. Ohba, Y. Kurauchi, T. Tago. "Real time manipulation of texture mapped surfaces". Computer Graphics, SIGGRAPH, 1897, pp. 181-188.
9. T. Kurihara, K. Arai. "A transformation method for modeling and animation of modeling and animation of human face from photographs". State of art in computer animation, Springer-Verlag, 1991, pp. 45-57.
10. M. Nahas, H. Huitric, M. Rioux, J. Domey. "Facial image synthesis using skin texture recording". Visual Computer, vol. 6(6), 1990, pp. 337-343.
11. K. Hattori, Y. Sato, "Handy Range Finder for Active Robot Vision", In Proc. of Int. Conf. on Robotics and Automation, May, 1995, pp. 1423-1428.
12. K. Hattori, Y. Sato, "Accurate Rangefinder with Laser Pattern Shifting", Proc. of ICPR, vol. C, 1996, pp. 849-853.
13. K. Hasegawa, K. Hattori, Y. Sato, "A Facial Measurement System for 3D Shape with Color Texture", The Journal of the Institute of Image Information and Television Engineers, Vol. 53, No.3, 1999.



eMoney Order System: The Smart way to Pay Online

Kasun De Zoysa¹, Rasika Dayarathna²

Department of Communication and Media Technologies,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7, Sri Lanka.

E-mail: ¹kasun@cmb.ac.lk, ²rasika@cmb.ac.lk

Abstract

This paper proposes a new payment system called eMoney order for the Sri Lanka Post by enhancing the existing money order system. It discusses the need for a new payment system for the developing countries with special attention in the Sri Lankan context. With e-commerce applications built using the proposed eMoney order system, there will be reasonable social and economic benefits for developing countries such as Sri Lanka. In this paper, the eMoney order concept as well as its social and economic benefits are discussed.

Keywords: Electronic Money Order, Electronic Payment, Sri Lanka Post, Credit Cards

1. Introduction

An electronic business cannot be conducted without having a “customer invisible payment system” such as a bank draft or a credit card. Most of the people who are in the developed countries carry out their business transactions using credit cards. However, very few people can afford a credit card in a developing country such as Sri Lanka since most of the people do not have a sufficient income level[1]. Researches in developed countries do not worry about such a situation since the percentage of people who cannot afford for the credit cards is negligible. As IT researchers in a developing country, we have to think of a new payment system, that would enable our people to participate actively in the Internet based financial transactions. eMoney order system is such a payment system which is build on top of the traditional money order system in Sri Lanka.

At present, the traditional money order system requires considerable time to transfer money and it cannot be used to pay over the Internet even though many improvements have been introduced to the system over the past years. Therefore the proposed

eMoney order system will improve the traditional money order system to meet today’s business and social needs.

2. Credit Card System

In the early days of Credit cards, they were used as a deferred method i.e. to pay at a later time for the goods and services presently rendered. Additionally, today, it is heavily used as a payment method on the Internet. Even though it inherits several weaknesses, it has been leveraged as a payment method on the Internet for quite a long time by now.

2.1 Mechanism

The mechanism of the proposed electronic money order system can easily be understood by studying the mechanism of the existing credit card system [2]. The existing credit card system works as follows:

1. The consumer gives his credit card to the merchant.
2. The merchant asks the acquiring bank an authorization code
3. The acquiring bank requests an authorized code from the issuing bank via the inter-bank network.
4. Confirmation is sent to the merchant bank from the acquiring bank.
5. The acquiring bank authorizes the merchant to proceed with the transaction.
6. The merchant proceeds with the customer’s order.
7. After a predefined period, the merchant presents a batch of sales details to the acquiring bank.
8. These details are forwarded to the issuing bank by the acquiring bank.
9. The issuing bank debits the customer’s account and sends the amount of money (after charging a service charge) to the merchant bank via the Inter-bank network.

10. The acquiring bank credits the merchant's account with the amount received from the issuing bank.

2.2 Weaknesses

Even though there are several weaknesses in the credit card system as mentioned below, it should be noted that the role played cannot be looked down upon since all other payment systems are based on this methodology [3].

1. **A single number** (Credit card number) is used for almost all the transactions carried out by the particular credit card. So once an eavesdropper gets that number he may use it for his transactions until the owner notifies the bank. In the proposed system, there will be a unique number generated for each transaction.
2. **Income level** – In Sri Lanka, to have a credit card, the income level of the applicant should be greater than Rs. 12, 000.00 per month. Monthly income and other requirements, which must be fulfilled to obtain a credit, card varies on type of the credit card and issuing bank. When we consider the income level of the Sri Lankan citizens, it is quite obvious that an average person may not be able to fulfill the most of these criteria to own a credit card facility.

The proposed system is designed to incorporate the plus points of the existing money a credit card can use the proposed eMoney order system.

3. Traditional Money Order System

Money order system was introduced to Sri Lanka by specially targeting the people who live in rural areas. In general, Figure 1 shows the present money order system [4]. However, improvements such as sending a money order via fax have been introduced into the money order system over the past few years.

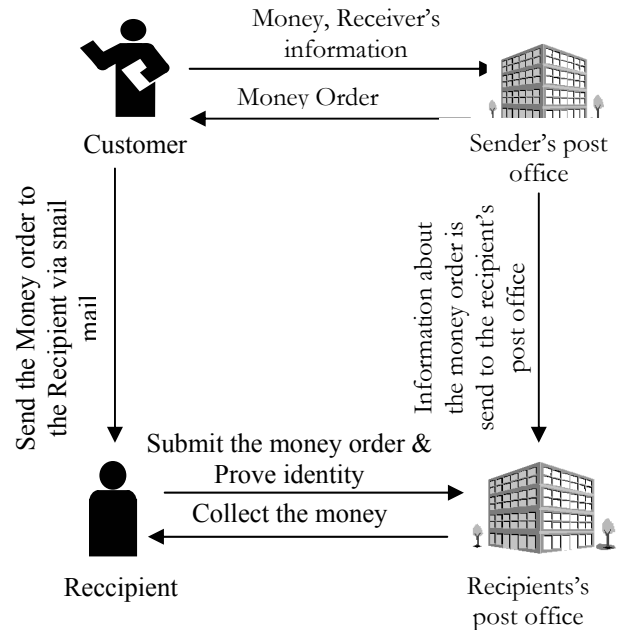


Figure 1: Traditional Money Order System

4. eMoney Order System

As in the present money order system, a user can purchase an eMoney order by paying the required amount and giving the following information.

- Recipient's information
- Sender's information

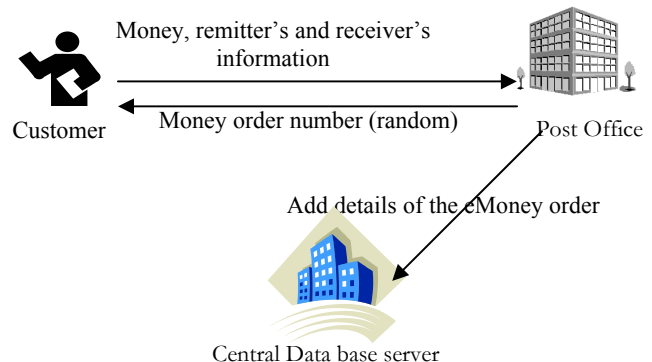


Figure 2: Purchasing a eMoney from a Post Office

A receipt corresponding to the eMoney order, which contains the eMoney order number, is issued to the customer by the post office and the following information is fed into the central database server by an officer at the post office (see Figure 2).

- Recipient's information
- Sender's information
- eMoney order number

The given money order number can be used as same manner in the existing credit card number to pay for an online transaction. This can only be done when the merchant supports this eMoney order system (see Figure 3).

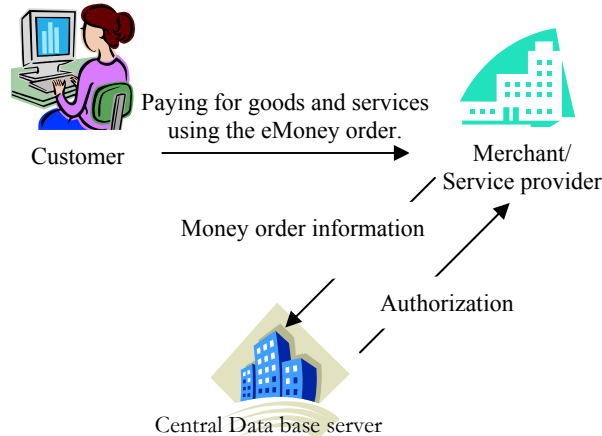


Figure 3: Using the eMoney Order

In general, once a recipient comes to realize an eMoney order, his authenticity must be verified. The recipient is asked to produce his identification number to verify the identity. The recipient may be an individual or a representative of an agency.

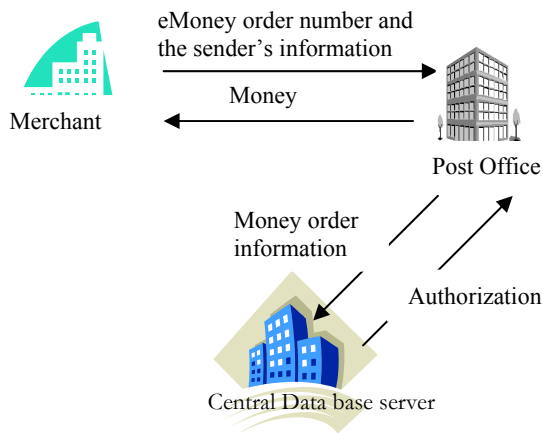


Figure 4: Collecting Money from Recipient's Post Office

Once the payment is made, validity of the eMoney order is checked against the central database server and this eMoney order is flagged as a used one before sending it to the recipient. Therefore, the recipient can make sure the validity of the eMoney order. As in the existing system, the purchased

eMoney order can only be used to pay for the original recipient.

Once the purchaser obtains an eMoney order, he cannot change the original recipient. This is a deviation from the credit card system where any amount can be paid for any recipient/merchant. The other important point is, if the purchaser applies for a service for which the identity of the purchaser is critical, then the particulars of the purchaser and the information of the applicant for that service has to be matched.

As in the existing system, the merchant/service provider can collect his dues by submitting the eMoney order number and the purchaser's information to his post office. That post office checks the validity of the order and purchaser's information before crediting the merchant's account (see Figure 4).

5. Systems Design

5.1 Architecture

The system is designed on the three-tier architecture consisting of the following sub modules.

Front end: Front end is a web interface which can be accessed using a web browser such as Internet Explorer, Netscape Navigator etc. In order to ensure efficient and secured access the latest version of the browsers or a services pack which supports 128 bit encryption must be made available at all post offices.

The Middle tier (application tier): Application tier acts as an intermediate module between the front and the back end and all the queries to the database go through this module. The business logic of the system is implemented in this module.

Back end: All the data are stored in the back end server. Oracle database is used since it provides higher security, efficient access and robustness.

Apart from the main database server, two other mirror database servers are used to overcome the single point of failure problem. Figure 5 shows the proposed system architecture. In this diagram, three databases are in one single location to represent the logical view of the system architecture, but physically these three databases should be located in three different sites. Each database is connected to the Internet via a firewall, which secures the data from viruses and other malicious attacks.

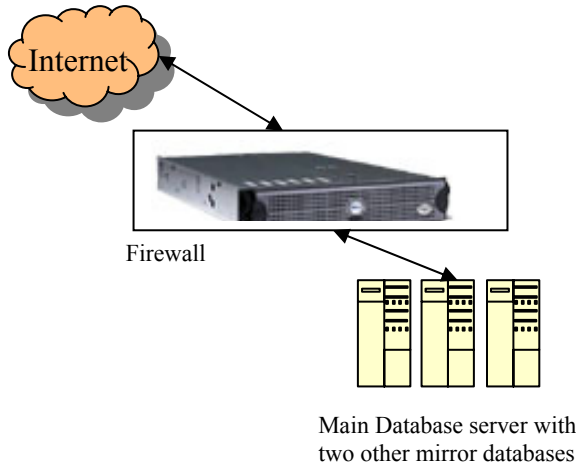


Figure 5: Proposed System Architecture

5.2 The System Roles

The operation of the system can basically be divided in to three parts based on the actors in the system.

- System users: Officer at a post office
- Customers: Regular money order customers, i.e., purchasers and recipients
- Administrator: A person who manages the entire system

System User: A postal officer can act as the system user in a post office.

Duties of the system user: As in the existing system a customer must fill an application named “*Application for Inland Ordinary Money Order*” with two additional pieces of information. Those are customer’s identification number and the recipient’s identification number. The existing money order application must be amended to incorporate these two numbers.

As already mentioned, an eMoney order is issued to the customer as in the conventional system. The additional work in the system is to feed the particulars of the eMoney order into the central database server. This can be done at the time of issuing the eMoney order (online) or at the end of the business day. If it is processed at the end of the day, the validation period should be started from the next day, otherwise it will not be considered as a valid one until the particulars of the eMoney order are available at the central database.

The other issue in online processing is that it is required to connect to the Internet quite frequently to update the database. To cope with this problem, the eMoney orders can be divided into two types as “*Express*” and “*Standard*”. An additional

commission has to be paid to buy an “Express eMoney order”. The additional cost for connecting to the Internet can be recovered from additional commissions. In case of a failure to work online, the additional commission can be refunded.

Customer: A customer can be a person or institution who purchases or receives the eMoney order. The following facilities are provided to him.

- Searching: Once an eMoney order is searched by using eMoney order number all the particulars of that eMoney order will be given.
- Submitting: Organizations such as University of Colombo School of Computing, Department of Examinations, Railway department etc which accept the payments through eMoney orders will have an entry for the eMoney order number in their applications. Filling the eMoney order number is almost similar to filling the credit card number in a web form. Instead of filling a web form, there are other ways of sending eMoney order number to a recipient such as through fax, telephone, email etc.

Administrator: An administrator’s tasks are as follows: accounting, creating, deleting and viewing the following accounts:

- System users
- Post offices

The proposed accounting system should produce all the required accounting reports in accordance with the standards followed by the Sri Lanka Post.

6. Conclusions

The eMoney order system enables the users who cannot afford credit cards to enjoy the same benefits presently available to the credit card users on the Internet. Without having any additional cost on the infrastructure, the existing post office network of the Sri Lanka Post can be used to implement the system. Since this is only an enhancement to the existing system, users can familiarize themselves with the system easily. The commission, which must be paid to the international credit card companies for each and every transaction, can be saved within the country. Since every transaction is carried out by using a distinct number, it is safer than the credit card system. Because of these reasons eMoney order system is well suited for developing countries such as Sri Lanka.

References

- [1] “Statistical Pocket Book”, Democratic Socialist Republic of Sri Lanka, Department of Census and Statistics, Ministry of Interior, Colombo, Sri Lanka, 2002
- [2] Jalal Feghhi, Jalli Feghhi and Peter Williams (1999), "Digital Certificates Applied Internet Security", Addison Wesley Longman, Inc.
- [3] Rolf Oppliger, "Security Technologies for the World Wide Web, Artech House Inc., 2000
- [4] Post Office Rules, 4th edition, Postal Department, Sri Lanka, August 1980



A Notarization Authority for the Next Generation of E-Mail Systems

Hiran Ekanayake, Kasun De Zoysa, Rasika Dayarathna

Department of Communication and Media Technologies,
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7, Sri Lanka.

E-mail: hbe@ucsc.cmb.ac.lk, kasun@cmb.ac.lk, rasika@cmb.ac.lk

Abstract

Current email system is at a risk of losing its demand because of its abuse. These abuses are ranging from receiving unsolicited emails to email frauds or repudiations. This paper discusses better approaches to overcome those limitations up to some extent with the concept of a trusted third party called email notarization authority.

Keywords: Secure Electronic Mail, Notarization Authority, Timestamping

1 Introduction

Electronic mail now has become the state of the art in distant communication. Its fancy is in its ease of use and cheapest cost. As a few years ago, one of the potential weaknesses with this service was its inability to provide a good security framework. But over the years there was a big effort to fix this problem and as a result now this service has a rich security infrastructure based on public key infrastructure [1].

However, email users are facing various other problems while using this service. There are lots of nuisance emails appearing on mailboxes daily. In other hand, emails have become the carrier for electronic viruses. Also once you get an email you cannot be assure that the actual sender of that email is the one who appearing on the “from” field of the email. Again, the current email system is unable to provide any legal proof for any party involved in an email transaction. As a result, current email system is losing its strength to provide a secure, reliable and trusted communication channel for today’s information exchanges [2].

Therefore, to stay with this service in the future, a more work has to be carried out in order to fix these weaknesses. Our new approach, which is based on a notarization authority, can solve most of the e-mail security issues relating to repudiation such as when an

email is created, who created it, when it was sent, was it delivered to the intended recipient, was it observed by the recipient.

2 Potential Solutions

There are different appearances for email systems: POP mail, and web mail. In addition, mail clients are differing as MIME [3] compatible, and S/MIME [4] compatible. Despite, email differs in various other ways: corporate mail, and other, all these together provide users the ability to exchange information with or without various contexts: security, reliability, ease of use, cost, etc.

Over the years several attempts have been made to increase the reliability of this service. For instance Outlook Express [5] allow read receipts to provide an indication of guaranteed delivery to its users. However, it has fallen to provide the requested service, because the recipient can ignore to send a receipt.

Under digital notarization concept, your email will be digitally timestamped, and later you will have a legal proof for your email transaction. These proofs vary from sending to reading of an email. ReadNotify [6] is a leading example in this area. There is no need to have client side plug-ins or any other modules to use ReadNotify notary service and they track the recipient(s) in a very transparent way. To receive the services from ReadNotify one has to register, and from that point onwards emails arrive by this address become eligible to receive the service. There are some other notary services, which enable similar functionality. However, none of them have not yet able to provide a comprehensive framework for strengthen the email system.

3 Role of the Notarization Authority (NA)

This section presents a discussion on how to expand the capabilities of the digital notarization to solve the repudiation issues with regard to email transactions. In order to solve these issues we proposed a trusted third party called “Notarization Authority” with the following capabilities.

- **Email Time stamping:** prevents backdating the existence of an email.
- **Proof-of-Posting Certificates:** provide legally acceptable evidence to prove that you actually posted an email.
- **Guaranteed Delivery:** will ensure that your email is delivered to the intended recipients; and after delivering you will be informed with a digital receipt.
- **Proof-of-Observing Certificates:** provide legally acceptable evidence to prove that your email is opened or read by the intended recipients.

3.1 Proof-of-Posting Certificate

This certificate proves that you have sent an email at particular time. Two different scenarios were used.

Scenario 1: The sender is a registered entity under the notarization authority. For each email send by the sender, a copy will be sent to that authority. Later the sender will receive a proof-of-posting certificate. Figure 1 illustrates this scenario.

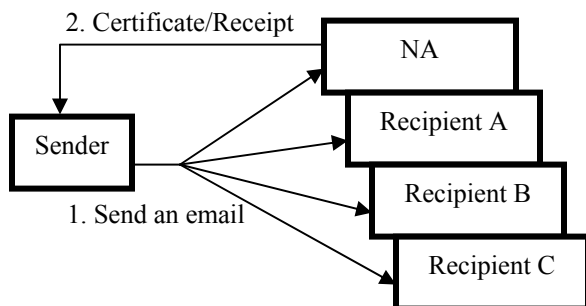


Figure 1: Proof-of-Posting Certificate Scenario 1

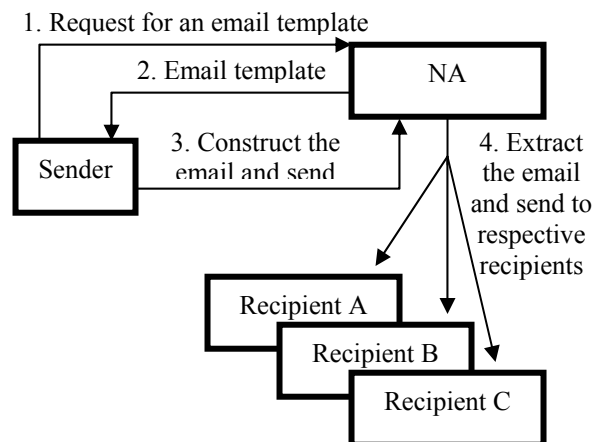


Figure 2: Proof-of-Posting Certificate Scenario 2

Scenario 2: The sender is a registered entity under a notarization authority. As the first step the sender requests a template from a notarization authority. Notarization authority responds by sending back a registered time stamped template. Based on this template the sender constructs his message, appends attachments, puts recipient email addresses, and finally sends it back to the authority. Notarization authority will then verify the template and extract the absolute email message from this filled template. This absolute message will be sent to the intended recipients. Figure 2 illustrates this scenario.

3.2 Proof-of-Observing Certificate

This certificate proves the actual observation of an email. Observation (reading) is somewhat difficult to capture. The following scenario describes how this event is captured using some existing technologies that works for both POP and web mails as well.

Scenario: The sender is a registered entity under a notarization authority. As the first step the sender sends email which need the proof to this authority. Notarization authority will then extracts the absolute message and encapsulates this message into a password protected zip file and send this email under a new envelope to those respective recipients with instructions on how to proceed. Each recipient has to go through these instructions, and has to visit the authority’s dedicated website to get the password to unzip the file back to the original email. This successful password-transferring event will be captured as the evidence of reading the email and a proof-of-observing certificate will be issued. Figure 3 illustrates this scenario.

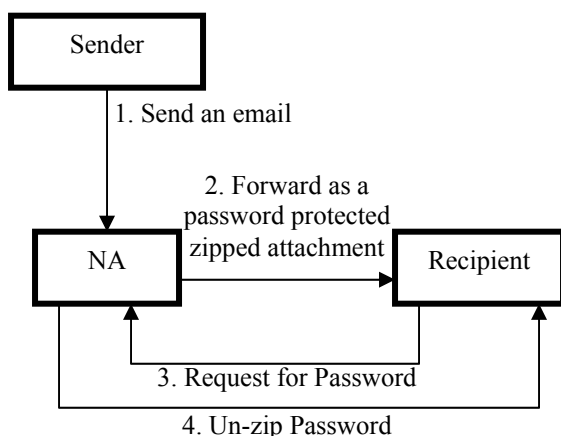


Figure 3: Proof of Observing Certificate Scenario

3.3 Spam Mail Prevention

With some few enhancements our notarization authority can be configured to filter Spam mails. Here it is assumed that all trusted email addresses are registered under the notarization authority. This authority has a mechanism to validate those email addresses and their respective owners periodically.

Scenario: The recipient is a registered entity under a notarization authority and this authority has a database of trusted email addresses. A filter protects the recipient's email client and configures to bypass emails only from trusted sources. All un-trustable emails will be forwarded to the notarization authority. The authority will then validate the respective sender against its trusted address database and forward back the emails with a report attached. Based on this report recipient's email filter may decide whether to drop it or not. Figure 4 illustrates this scenario.

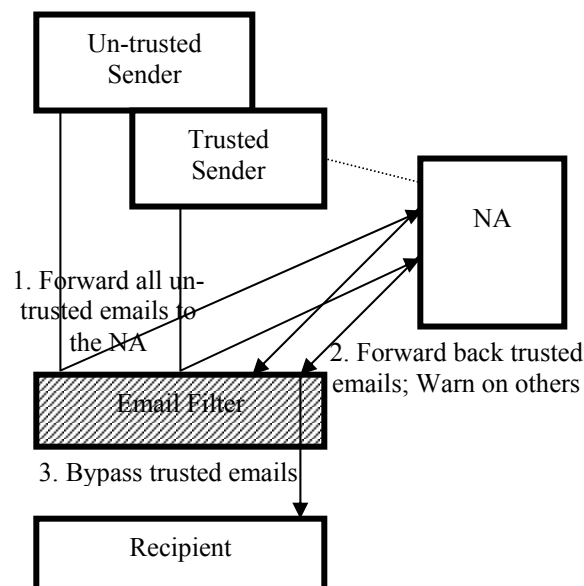


Figure 4: Spam Mail Prevention Scenario

4 The System Architecture

This section describes the system architecture with the design aspects. As illustrated in Figure 5 the system is divided into two sub-systems:

- **The Web-based Service:** provides services to the users
- **Notary Server:** executes periodic functions and security intensive functions

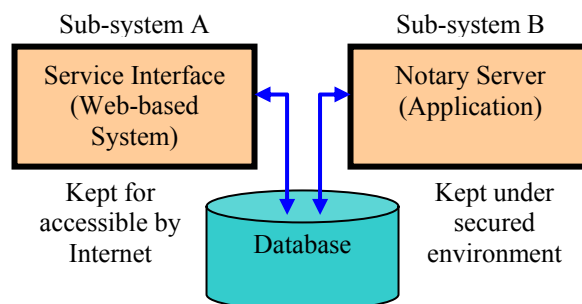


Figure 5: Sub-Systems of the Main System

The database acts as a common gateway to these two sub-systems and separates the two into two different functional domains.

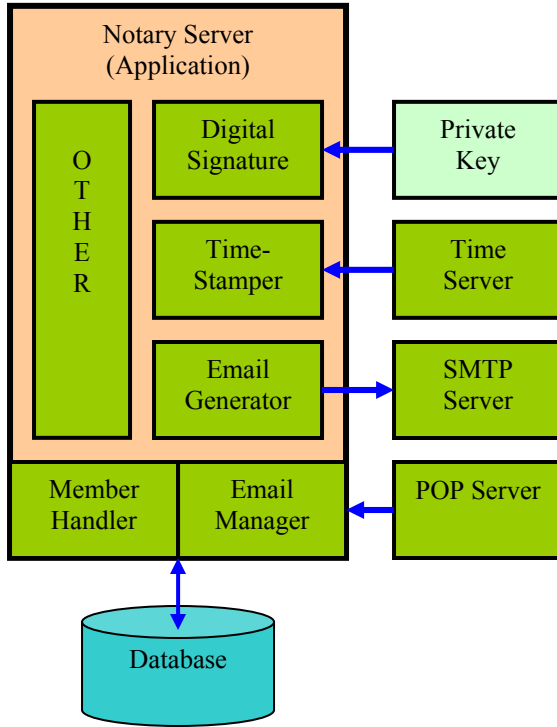


Figure 6: Notary Server Sub-system

The notary server is the hart of the system. It is to be kept under a secure environment, and its main purpose is to execute periodic functions that need some higher degree of security. These functions includes:

- **Email Handling:** both sending and retrieving
- **Time-Stamping:** synchronizes the time with a remote time server and provides time stamping service to certificates and receipts.
- **Issuing Proof-of Certificates:** construct digitally signed proof-of-posting and proof-of-observing certificates for those applicants.
- **Member Registration:** validates new member registrations and periodic verification of information.
- **Message Handling:** feature extraction from email messages, enveloping and email constructions.

Figure 6 illustrates components of the notary server sub-system where these functions are executed.

The Web service sub system offers the following services:

- **Web-mail & Message Templates:** web mail offers the facility to construct emails without using a separate mail client. Message template differs according to the purpose of the sender. For instance, if you need to send a birthday greeting email on a pre-determined day you have to use a greeting message template.

- **Email Status:** offers the facility to query the current status of an email.
- **Certificate Issuing Facility:** takes the information from requester and sends a proof-of certificate.
- **Certificate validation facility:** validates any issued certificate.

Figure 7 illustrates components of the web service sub-system where these functions are executed.

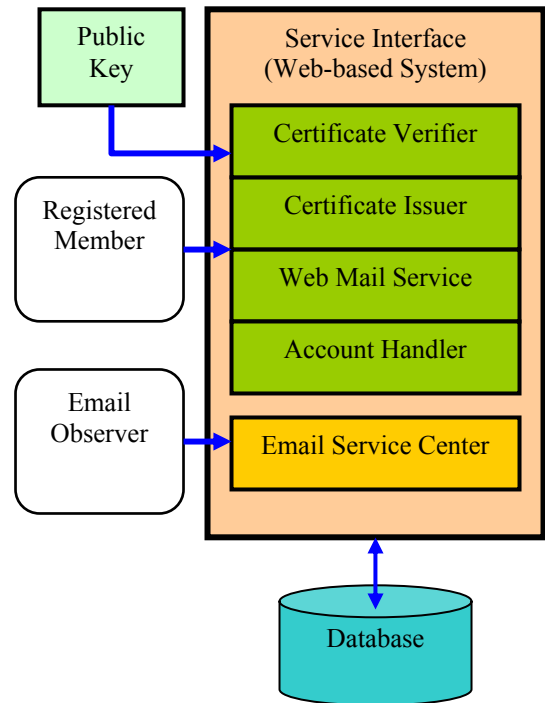


Figure 7: Web Service Sub-System

5 Conclusions and Future Works

5.1 Conclusions

This paper described some methods to make the existing email system a much perfect one based on existing technologies and standards. Several initiators have already built notarization authorities for emails using their own approaches. However none of them provide a complete solution. This is because current email protocols such as SMTP and POP [7] do not provide significant messages for such enhancements.

However, our proposed notarization authority enriches in facilitating non-repudiation and Spam mail filtering. These methods will be useful for designing a better email system for the next generation.

5.2 Future Works

Email tracing facility is not incorporated into the system yet. If the tracing facility required the sender could get the status information on how an traveling with the location and time information. So he can visualize the current residence of any email.

Some extensions to the email, for instance blocking or forwarding, are very easy to implement in the system. For example, the email forwarding can be done in the following manner:

- **Direct forwarding:** forwards an email to someone else other than the recipients mentioned in the email.
- **Cluster forwarding:** forwards single email to a group of recipients.
- **Chaining:** If the email can't reach to the recipient A then forward it to recipient B; if that attempt also failed then forward it to recipient C, and so on.

References

- [1] "Cryptographic Message Syntax Standard", Public-Key Cryptography Standards, RSA Laboratories, [Online] Available at <http://www.rsasecurity.com/rsalabs/pkcs/pkcs-7/index.html>
- [2] Rasika Dayaratna, Kasun De Zoysa, "An Enhanced Security Mechanism for General Purpose Email Clients", International conference on Computer Communication, The International Council for Computer Communication, Mumbai, India, August 11-14, 2002
- [3] "MIME related links", Multipurpose Internet Mail Extensions MIME, [Online] Available at <http://www.oac.uci.edu/indiv/ehood/MIME/MIME.html>
- [4] "S/MIME Version 3 Message Specification", Request for Comments, [Online] Available at <ftp://ftp.ietf.org/rfc/rfc2633.txt>
- [5] Microsoft Outlook Express, Microsoft cooperation, [Online] Available at www.microsoft.com
- [6] ReadNotify notary service, [Online] Available at <http://www.readnotify.com>
- [7] Rolf Oppliger, "Security Technologies for the World Wide Web, Artech House Inc., 2000



Learning Patterns: Towards the Personalization of E-Learning

Dr. K. P. Hewagamage and R. S. Lekamarachchi
University of Colombo School of Computing (UCSC), Colombo, Sri Lanka
E-mail: kph@ucsc.cmb.ac.lk and surangga@yahoo.com

Abstract

E-learning through the web began as a service, by publishing some educational materials at websites. At the early stages, the learning process was carried out by browsing such a collection of educational materials and its effectiveness was based on the presentation of content in those materials. Customizing the presentation and interactivity of an e-learning application based on the users (which we call here personalization), should be done based on each learner's knowledge and learning experience. A learning pattern, presented in this paper, is a meta-level piece of information that can be used to achieve low coupling personalization in an e-learning courseware. A learning pattern, which is modeled on different levels of abstraction, is constructed based on the interaction of a learner with relevant materials and it is depreciated when he/she stops accessing them. In this paper, we also present how a learning pattern can be used to identify the appropriate learning path in an e-learning courseware. A collection of learning patterns is used to describe one's learning experience.

1. Introduction

Learning is the only thing next to breathing that we do continuously from the day we were born to the day we will die. By everything you do, say or think, you learn something. People have different definitions and views on learning but ultimately you will have to end up with a common phrase, “*You cannot live a single moment without learning something.*”

The methods of learning have evolved over time and they have come a long journey from the exercise book to the palm top computer with the advancement of technology. However, the basics still remain the same.

E-learning is a process of learning that takes place through a network, usually over the Internet or an intranet of an educational institute or a company. It has its roots in the not-so-attractive world of computer-based training (CBT), which appeared in the early '80s

and used CD-ROMs to teach mostly technical skills to technical people. Lately, e-learning has evolved into a tool widely used in both the corporate and academic worlds. With the rapid expansion of the World Wide Web and other Internet based services, e-learning is becoming a tool for everyone's lifelong learning.

Learning is a highly personalized activity which varies based on the background knowledge of the subject, previous experiences (general and specific), motivation, preferences, context and other activities in which the user is engaged. Personalization is a process which would affect the learning of an individual at different levels [11]. First, it should identify specific skills and knowledge gaps and direct learners to the appropriate lessons or modules. Once a particular learning activity is started, it is possible to customize the presentation of learning materials according to factors given earlier in order to maintain an effective, efficient and continuous process of knowledge transformation. When the learner undergoes this change due to the knowledge acquisition, the values of all influencing factors of learning activity are also modified and such variations should be communicated to maintain an effective personalization process.

In this paper, the authors discuss a metadata based description which they name as a learning pattern, and how it can be built and utilized to achieve low coupling personalization while protecting the learner's privacy. In many traditional applications, personalization is considered as a tightly bound process which undertakes monitoring the user's activity and maintaining the user's profile. But if the learning service is provided by an unknown third party the learner will probably refuse to provide personal information that is simply needed for the personalized service described earlier.

Generally, the personalization is provided as an interface service to the user in a circular process. The success of this service depends on the corporation of three parties namely, the learner, author and facilitator. The author, who defines the sequencing order of lessons, should include enough metadata with each lesson in order to provide multiple sequences of presentation. The facilitator, who is the publisher of the copyrighted learning materials, will have to interpret

the learner's learning pattern in order to decide the most effective sequence of the presentation in an e-learning course.

This paper is organized as follows. Section two describes issues with respect to learning content and packaging. In section three, we define learning patterns and section four illustrates how learning patterns can be used to realize personalization in an e-learning package. Section five briefly covers about an intermediate website "Learning Home" to maintain learning patterns. We discuss details of relevant standards and related user models for personalization in section six. Finally, we conclude the paper stating benefits, limitations and the future of learning patterns in section seven. There are also several diagrams used to illustrate some of the important facts described in this paper.

2. Learning Content and Packaging

If the content of a particular entity/object is supposed to provide learning to the people who will access or interact with that, it should provide information that could generate new knowledge in the learner's mind. However, this process is more complicated than what we anticipate, since the process of transforming information into knowledge heavily depends on the learner's personal status and the form of communication.

The learning content in a particular courseware should be structured according to learning objectives specified at the very beginning of the course [10]. Number of modules in a courseware is determined based on these overall objectives and each module is specified with number of sub-objectives. The advantage of these objectives is that the author can modify the courseware to add/delete new content when he/she wants to create a different version of the courseware for a new offering. Based on the sub-objectives of each module, the relevant number of lessons is determined. Generally, a sequence should be specified when lessons in a module are combined.

According to the instructional design principles [12], a lesson should not communicate many ideas in a single visual display. A single visual display is referred as a page in the e-learning courseware and a lesson may consist of one or more pages depending on what extent the author wants to present the relevant materials to the learner.

2.1 Learning Styles

A learning style is a mechanism to deliver a particular content in a lesson, based on the user's preferences and/or the most effective way to do so. This can be

based on the learner's background knowledge and skills. In some other works, the personalization is discussed as the selection of a suitable learning style [12]. However, personalization should go beyond the mere selection of a suitable learning style.

Identifying and categorizing learning styles into a common set of styles have been done by a number of studies. The process of personalization can be started by selecting an appropriate learning style in a lesson. Following four categories of learning styles are widely used to design the presentation of e-learning courseware.

Style	Description
Visual/Verbal	Prefers to read information
Visual/Nonverbal	Prefers graphics or diagrams to represent information
Auditory/Verbal	Prefers to listen to information
Tactile/Kinesthetic	Prefers physical hands-on experiences

2.2 Learning Path

The structuring of learning content based on different levels of objectives, which is known as packaging is usually done by the course author. A package, which is the technical representation of an e-learning course, usually contains additional information about how amalgamated learning content should be sequenced in a presentation and some additional metadata to describe related learning resources. The learning path is a specific sequence of learning content [9][10]. Hence, a package may contain a number of learning paths depending on the author's metadata specification.

Figure 1 shows a simple course structure in a package. It has a number of learning paths and the browser which presents the e-learning courses, will select a relevant path depending on the meta information given in the learner's profile and his/her performance at different quizzes.

Sometimes, a particular learning path could be considered as a version of the course defined for a specific user group while still supporting common learning objectives. A simple knowledge pre-test could be used to determine which learning path is the most suitable at the beginning.

The size and scope of learning content that are combined to form an e-learning course is also a key consideration. For example, if a package is comprised

of only a few learning assets, then it may not make any sense to define different learning paths. This issue is also very important in the reuse of learning content [6].

3.0 Learning Patterns for Personalization

When you go to your shoe dealer you try to find a pair of shoes that perfectly matches your needs. The shoe manufacturer produces shoes in different sizes, colors and styles so that people with different preferences can find their matches. You select the best matching pair for you. Sometimes this best match is not the perfect match based on your requirements. We can think of the personalization as the perfect match for you. However, achieving the personalization is not easy as one would imagine. Therefore, when there is no alternative, you are forced to accept what seems to be the best match.

A pattern is an abstract representation of knowledge in a particular context. It could be used to document a group of entities/objects at different levels of abstraction. At the same time, it could be used to identify a solution in a particular context. A learning pattern corresponds to such a representation with respect to an individual in the domain of his/her e-learning space. A learning pattern which has an XML type representation, carries all information required for the personalization of the learning experience.

Since learning, which takes place with the interaction in the e-learning space, modifies the learner's knowledge, the corresponding learning pattern will also grow. This learning pattern is different from the static representation of the user's profile which just gathers metadata with respect to a pre-defined structure.

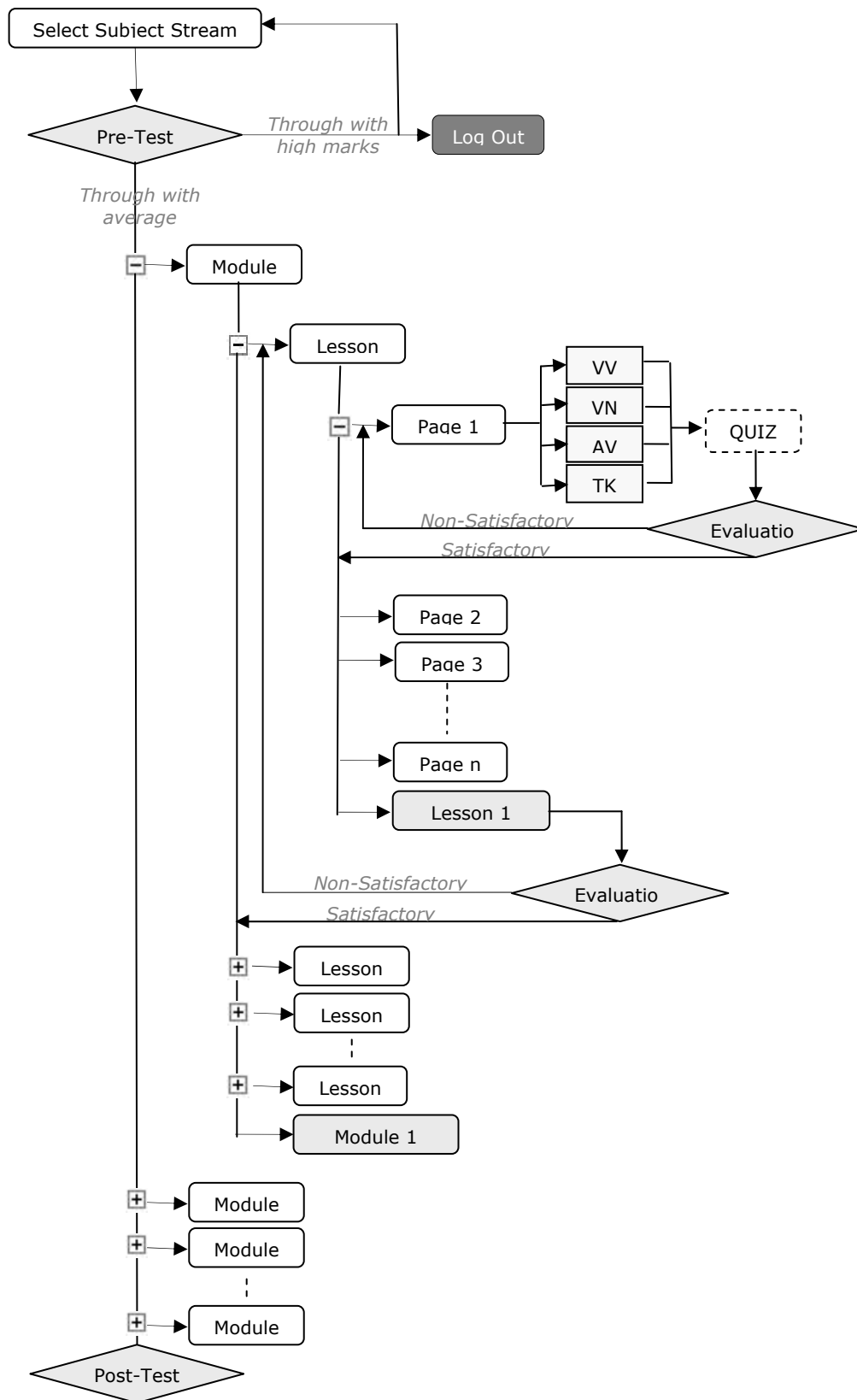


Figure 1: The course structure in an e-learning package.

At the same time, it could provide information at different levels when required. In addition, the learning experience is interpreted to specify its content.

A single user may have a lot of learning patterns with respect to the stream of subjects he/she is learning. This categorization can be done with respect to the scope of subject or classified interests of the learner. Hence, a learning pattern would be initiated whenever he/she takes the first course in a classified area of study/interest. The level of abstraction in learning patterns will be high at one end and it gives more details at other end.

Whenever the learner takes a related course in those classified areas, the corresponding learning pattern will be updated. But, with time it will be depreciated, since people usually forget what they know with the pass of time. Generally, someone's learning experience can be described as a collection of such learning patterns organized in a triangular model as shown in Figure 2. It divides a learning pattern into three meaningful sections.

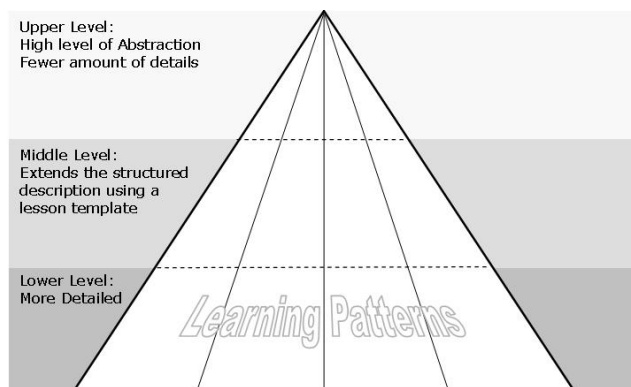


Figure 2: *The triangle represents a collection of learning patterns to show one's learning experience.*

A learning pattern has three levels, namely upper, middle and lower levels. In the upper level, the abstraction is very high and it describes a particular learning experience in a summarized way. The middle level of a learning pattern extends the previous abstract definition using a structured description. The lower level organizes the user's learning experience captured based on his/her active learning [8] activities.

3.1 Upper Level

The upper level includes some abstract details about the subject area using keywords and phrases. Usually main titles and other sub headings of the courses followed are used as phrases in the upper level. Author specified metadata or a sampling technique [7] could be used to identify relevant keywords. The upper level also defines the user's learning experience using a statistical counter (tag named 'success factor') and a classification based on the possible modules. When a learning pattern is initiated, modules are inherited from the first course but later they are split, merged or expanded based on the user's interaction with related other e-learning packages.

When a learner follows a course, he/she has to take multiple choice questions or other interactive assignments for which he will be given marks. The success factor is calculated by amalgamating these marks using a weighted average scheme with respect to the whole course as well as in module level.

3.2 Middle Level

The middle level further extends the description of modules given in the upper level. A module is described as a collection of lessons. Then a number of lessons simply specifies the scope of a particular module and it is a useful heuristic when two packages are compared. A lesson is defined using keywords, phrases and success factors.

3.3 Lower Level

This is the level which describes the user's interaction with respect to each lesson given in the upper level. A single lesson could be presented in different styles (Section 2.1) in order to provide the most suitable way for each individual learner to absorb the content of the lesson. Passive reading, watching and listening could hardly deposit a new knowledge in a learner's mind. The success factor is a good quantitative measure to evaluate the level of absorption but it doesn't depict the knowledge structure. The actual knowledge structure in a learner's mind is not visible but the user's activities such as underlining, highlighting and commenting could be used to approximate the outline of a structure. Some other applications can use the lower level to provide a quick revision to refresh the learning experience.

Figure 3 illustrates a fraction of the learning pattern for a course in mathematics. Such a pattern will be depreciated with time if the learner does not engage in the learning activities of the corresponding subject.

4. Realizing e-Learning Personalization

In the previous section, we discussed how to build learning patterns and their technical infrastructure. As we have described, these learning patterns are used to provide low coupling based personalization. In this paper, we are only discussing how these learning

patterns could be used to provide a customized learning sequence.

By matching relevant keywords and phrases, the system that facilitates browsing e-learning packages, first tries to identify whether a relevant learning pattern exists whenever the user selects a new course. If such a pattern is identified, it calculates a depreciation considering the date the learning pattern was created, the last date pattern was updated, and the current date of the system.

$$\text{Depreciation} = \frac{\text{LastDate} - \text{CreateDate}}{\text{CurrentDate} - \text{CreateDate}}$$

```
<?xml version="1.0" encoding="ISO8859-1" standalone="yes"?>
<abstract level>
<title phrases>
  <phrase> University 1st year Mathematics </phrase>
  <phrase> e-maths, online mathematics home page </phrase>
  <phrase> Applied Mathematics </phrase>
  .....
<keywords definition>
  <keyword> Differential Equations </keyword>
  <keyword> Integration </keyword>
  <keyword> Statistics </keyword>
  .....
<keywords definition>
<meta details>
  <user status> satisfactory </user status>
  <last visited> 12-oct-2003</last visited>
  .....
</meta details>
<success factor> 65% </success factor>
<module structure>
  <module 1>
    <title> Motion </title>
    <keywords definition>
      <keyword> Motion in a Straight Line </keyword>
      <keyword> Motion in a Circle </keyword>
      .....
    </keywords definition>
    <success factor> 77% </success factor>
    <number of lessons>10</number of lessons>
  </module 1>
  <module 2>.....</module2>
  .....
  <module n>.....</module n>
</module structure>
</abstract level>
<middle level>
  <module 1>
    <lesson 1>
      <title> Motion in a Straight Line </title>
      <keywords definition>
        <keyword> Velocity </keyword>
        <keyword> Acceleration </keyword>
        <keyword> Distance </keyword>
        <keyword> Height </keyword>
      </keywords definition>
      <success factor> 90% </success factor>
    </lesson 1>
    <lesson 2> ..... </lesson 2>
    .....
    <lesson 10> .....</lesson 2>
```

```

</module 1>
<module 2>.....</module 2>
.....
<module n>.....</module n>
</middle level>
</lower level>
  <module 1>
    <lesson 1>
      <page 1>
        <learning style 1>
          <time> 3:35:05 </time>
          <sucess factor> 55% </sucess factor>
        </learning style 1>
        <interactivity>
          <underline> .....</underline>
          <highlighted> .....</highlighted>
        </interactivity>
      </page 1>
    </lesson 1>
    .....
    <lesson 10> ..... </lesson 10>
  </module 1>
  .....
  <module n> ..... </module 10>
</lower level>

```

Figure 3: A sample learning pattern for a course in Mathematics.

The depreciation value can be used to approximate the current value of the success factor. Rules are defined for different levels of personalization based on the current value of success factor. The common rule is that if the value is less than 45, then there will be no personalization of learning sequence in the new e-learning course. If it is less than 75 and greater than or equal to 45, then a medium level personalization will be considered. A higher level personalization is applied, if it is greater than 75. These common rules can be modified according to the learner's preferences by specifying different levels of personalization.

The next step is to identify matching modules in the learning pattern and e-learning courseware. This is done by matching the most relevant keywords and phrases used to define each module in the learning pattern with modules in the e-learning package. If a module in the package can be matched with a module in the pattern, the scope of these two modules are compared considering the number of lessons. Hence, three possibilities are identified "less than", "equal" and "larger than". The success factor is calculated considering the depreciation of the pattern and the recorded success factor for the module. Rules for the sequence modification are as follows.

Module Current Success Factor (msf) = Depreciation * (Recorded Success Factor for the module)

IF ($msf < 45\%$) **THEN** {no_sequence_modification}

IF ($msf \geq 45\%$ **AND** $msf < 75\%$) **THEN**
 { **CASE OF** "ModuleScope"
 "less": learning path discarded in whole module
 "equal" **OR** "large": obtain user's direct response }

IF ($msf \geq 75\%$) **THEN**
 { **CASE OF** "ModuleScope"
 "less" **OR** "equal": learning path discarded
 "large": only advanced lessons considered }

In the comparison of keywords and phrases declared between modules in the pattern and modules in the package, if there is a mismatch, then it is investigated considering the following possible cases. In case of such a mismatch, the module is considered to be a new module and is inserted into the pattern. If only a subset of keywords is matched, then it would be considered as a module with the less scope and the same rules will be applied to determine its personalization in the learning path. On the other hand, if it is verified that the scope is covered by two modules in the pattern, then the personalization is carried out at the lesson levels.

5. “LEARNING HOME”: The Place to Maintain Learning Experience

As we know, there are many websites which host e-learning courses on both paid and unpaid basis, and the number is increasing daily. Most of the packages available in those services are not integrated with a learning management system (LMS). Hence, it is sometimes hard to provide personalization facilities since the user is not willing to share his/her learning experience with a third party just for an additional facility.

As a solution, we propose an intermediate website which acts like the learning portal of the user. All learning patterns, which belong to a particular user, are stored at this website called *Learning Home*. It provides an interactive interface for all these patterns and when it is required, the user can also open and edit any learning patterns.

In a prototype system, the browser retrieves information from Learning Home to identify the correct pattern when the user opens a learning package. The system can also update learning patterns with new information or can insert new patterns when the browser interprets the current package as a new course. This pattern is updated as and when the learner takes lessons from this course.

6. Related Standards and Work

As we mentioned earlier, the personalization can be realized only through the cooperation of three parties; the learner, author and facilitator. In order to achieve such cooperation, e-learning standards play an important role.

IMS Content Packaging [1] is an interoperability specification to allow content creation tools, learning management systems and run-time environments to share content in a standardized set of structures. The purpose of this specification is to provide a mechanism that will allow content to be exported between systems with minimum effort.

In 1997, US Government initiated to push development plans for standardization of learning resources. As a result, Sharable Content Object Reference Model (SCORM) [2] was established by linking a number of other standards. The ADL SCORM constitutes three key components:

- The Content Aggregation Model (CAM),
- The Run-Time Environment (RTE), and
- Content Packaging.

Through these components the ADL aims to meet the following high-level requirements.

- **Reusability:** the content persistence over different LEs using a unified method of content markup.
- **Accessibility:** the globally accessible content repositories using metadata search facilities.
- **Durability:** the persistence of learning resources and system components over time.
- **Interoperability:** the platform independence of learning systems and learning resources.

“Learning Pattern” is a concept defined by combining interaction patterns and user profiles. There are some other similar models that have been described in the literature [3] [4] but learning patterns use the abstraction to model the complexity and carry the user’s manipulation interaction to provide personalization in different learning environments.

Creating fixed stereotypes is one of simple ways of user modeling [3]. New students are categorized and the system will customize its performance based on the category that has been set for each student. For example users could be categorized into novice, intermediate and expert levels within a system. This approach is useful when a quick, but not necessarily accurate assessment of the user’s background knowledge is required [4].

The overlay model is widely used in the adaptive hypermedia systems in the educational domain. A model of the student’s knowledge is constructed on a concept-by-concept basis and updated as the user progresses through the system. This allows a flexible model of the student’s knowledge for each topic [5]. For this model, the knowledge domain must be modularized into specific topics or concepts, similar to learning pattern concept given in this paper.

7. Conclusion

In this paper, authors presented a model for personalization, named learning patterns, for e-learning courses. This model is defined by extending the conventional approach of user-profile which is usually used in personalization of many applications, by especially considering the structured learning process.

Learning is a highly varying process from person to person depending on the individual skills and abilities. As a result of this process, the learner’s level of knowledge grows giving him more power to interact with a learning environment. We introduced the learning pattern concept to depict such changes in learning environment. A learning pattern is supposed to grow, freeze or die with time depending on the learner’s interaction with corresponding learning materials.

A learning pattern corresponds to a particular subject or interest in a learner's learning space. It is initiated when the learner starts the first e-learning course in a particular subject or interest. A learning pattern is updated every time the learner follows a similar learning content. In the scope of the work discussed in this paper, we only presented how it could be utilized to customize the learning path which is defined by the sequence of modules, lessons and pages in an e-learning courseware. However, it has more potential to provide fully fledged personalization. For example, it is possible to customize the visual appearance in a page using information recorded in the lower level of a learning pattern.

A learning pattern is depreciated with time when there is no interaction with relevant materials. This value is used to determine the learning sequence while the user interacts with corresponding packages. This sequence could vary in the same package at different times since the calculated values could be different.

One of the main limitations of the approach is that these learning patterns heavily depend on the keyword matching algorithms and metadata provided by the author. A learning pattern is initiated when the user takes his/her first relevant course. If the first package doesn't provide a suitable structure, it could badly affect the maintenance of the pattern. In this case, a learner can modify the pattern by accessing it at the "Learning Home" which maintains his/her learning experience using these patterns.

In the future, we will work to extend the functionality of learning patterns to provide different types of personalization. We emphasize personalization as not only providing what the user wants but also providing it just in the way he/she wants. Hence, we hope to integrate active learning and learning patterns while adhering to immersing standards for e-learning such as ADL SCORM.

Acknowledgement

The authors would like to convey their sincere gratitude to the staff of University of Colombo School of Computing (UCSC) for their constructive comments and suggestions.

References:

- [1] IMS CP (IMS Content Packaging), Public Draft, Version 1.1, December 2000.
- [2] ADL. "ADL Sharable Content Object Reference Model", Version 1.3 , <http://www.adlnet.org/> [10th Nov. 2003]

- [3] Rich, E "Stereotypes and User Modeling", in *User Models in Dialog Systems* pp. 35-51. Springer, Berlin, Heidelberg, 1989.
- [4] Kobsa, A. "User modeling: Recent Work, Prospects and Hazards" in *Adaptive User Interfaces: Principles and Practice*, M. Schneider-Hufschmidt, T. Kühme, and U. Malinowski, (eds.). 1993, North-Holland: Amsterdam.
- [5] Brusilovsky, P. "Methods and techniques of adaptive hypermedia" in P. Brusilovsky and J. Vassileva (eds.), *Spec. Issue On Adaptive Hypertext and Hypermedia, User Modeling and User Adapted Interaction 6 (2-3)*, pp.87-129, 1996.
- [6] Conlant, O., Wade, V. Bruen, C. Gargan, M., "Multi-Model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning" in *proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2002)*, April 2002.
- [7] C. Jayawardana, K. P. Hewagamage and M. Hirakawa, "Personalized Information Environment for Digital Libraries", *Journal of Information Technology and Libraries* (published by American Library Association), Vol. 20(4), December 2001.
- [8] C. Jayawardana, K. P. Hewagamage and M. Hirakawa, "Personalization Tools for Active Learning in Digital Libraries", *The Journal of Academic Media Librarianship*, Vol. 8(1), Summer 2001.
- [9] "Courses Personalization in E-learning Environment" V. Carchiolo et. al. in *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies*, July 2003.
- [10] "Courses and Exercise Sequencing using Metadata in Adaptive Hypermedia Learning Systems", S. Fisher in *ACM Journal of Educational Resources in Computing*, Vol 1 (4), Spring 2001.
- [11] "Agent Support for Personalized Learning Service" T. Hawryszkiewicz, in *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies*, July 2003.
- [12] William J. Rothwell, *The Fundamentals of Instructional Design*, University Park campus of The Pennsylvania State University.



The Effectiveness of Digital Government as a Tool for Improved Service Delivery

Mehdi Asgarkhani

Faculty of Commerce, C P I T, Christchurch, New Zealand

Abstract

Today, as the pace of change accelerates, access to information and communication technologies (ICTs) is critical for economic and social development. Changes include the rapid expansion of the Internet and the network economy; increasing business and individual use of wireless communications; access devices and new ICTs; the convergence of technologies, services and markets; globalisation of markets and trade; greater competition; mergers and acquisitions and their impacts on the restructuring of industry and markets. These changes can potentially transform all aspects of society, work, business, and government. The Internet and Web-based technologies have both had a profound effect on the way(s) in which the public sector functions - in that it has made it possible for many innovative government agencies to think of new ways in which to use the Internet in order to provide Web-based services to citizens. This paper examines the effectiveness of digital government solutions through: reviewing the views and perceptions with regards to the implications of digital government; elaborating on the 'digital divide' and its impact on the success of digital government; and outlining the results of a pilot study of how citizens viewed digital government initiatives.

1. Introduction

Over the past few years we have witnessed rapid advancements in ICT – which has in turn contributed towards a staggering growth in global computer networking and the emergence of a globally connected world. Increasing adoption of ICT has changed the way in which many businesses and organizations operate. The Internet has evolved from being a network for researchers and academics into a platform that has enabled new businesses to find alternative ways in which to offer their products and services. Web-based technologies and the Internet have fundamentally transformed the technological, economical and social landscapes. Within the so-called information society, critical information, disseminated across global networks, can radically transform the balance of power

among institutions, governments, policy makers, and people.

As globalisation of national economies intensify competition, much of the private sector has had to consider a stronger commitment to customer service as being critical for survival within the highly competitive business environment. Consequently, the private sector has increasingly set high standards of service both domestically and internationally. Government departments and agencies, however, were initially slow to respond to the challenge alongside the private sector.

As public awareness of potential benefits of the Internet and Web-based solutions continues to grow, there is an increasing expectation that government agencies would provide the same level of service as the private sector - therefore placing pressure on governing bodies to incorporate Internet resources into traditional governance practices. Over the past few years the use of ICT to enhance public sector management practices, known as “electronic or digital governance” has become increasingly popular. We are witnessing an increasing number of government institutions and agencies (local, federal and/or national) attempting to offer a wide variety of Web-based solutions and services. Politicians across the world focus much of their strategies on modernising government through adopting new technologies [9]. The search for a better government (worldwide) is a global preoccupation. Recent developments in digital government are not limited to developed countries. There are numerous examples of the introduction of digital government solution in developing countries. For instance, the government of Sri Lanka responded to recent developments through launching a vision for e-Sri Lanka - which outlines a strategy for the ways in which ICT can be employed in order to support the new wave of social and economic developments.

The introduction of electronic or digital governance practices causes a paradigm shift away from traditional information monopolies within governments. Governance practices have traditionally operated on a hierarchical model of information flow and interaction. In this model, information flows from a single governmental source through a system of designated recipients, where it is passively received and acted upon. In the upward flow of information, the feedback from the society flows through a limited number of channels, and is integrated into a centralized governmental source. For

example, information flows from elected officials through to public servants operating within fixed departmental structure and from there onto passive citizens. Limited feedback is provided in the form of elections. In addition, boundaries to horizontal information flow through the government's departmental structure often exist, as well as boundaries that hinder citizens' ability to evaluate the effectiveness of policies. The introduction of digital governance reduces these traditional hierarchies in governmental practices and creates an environment where information flow is bi-directional.

This hierarchical model is generally applicable to governance and public administration practices (see example mentioned above.) Little interaction or upward flow of information takes place between elections. In addition to this, boundaries to horizontal information flow through the government's departmental structure often exist, as well as boundaries that hinder a citizen's ability to evaluate the effectiveness of policies.

Overall, it is fair to say that the challenges to effective governance in a knowledge society are profound. In many areas, we are faced with new questions with as yet unknown answers. Over the past few years, there has been much debate over the success and usefulness of digital government. There is some evidence of a reactive approach when adopting digital government solutions. Consequently, some countries may end up with solutions that do not appear to be compatible with their specific needs.

This paper elaborates on the effectiveness (success and usefulness) of digital government. Three different facets of effectiveness have been examined:

- Effectiveness as the view of management and ICT strategists - with regards to the implications of digital government.
- Effectiveness as the implications of digital divide and e-readiness - with regards to digital government.
- Effectiveness - as the customers'/citizens' view of the usefulness and success of digital government.

2. An Overview of Digital Government

In this paper, the term 'government' (as act of governing) does not necessarily refer to 'governance.' Governance is viewed as a process - a 'guiding process' through which companies, organisations, groups and societies make decisions and manage their day-to-day activities and the ways in which they interact with one another.

Digital or Electronic Governance (E-governance) is a term used for emphasizing the application of Information and Communications Technology (ICT) in governance systems and processes – in order to adjust the theory and practice of legitimate decision making and policy formulation so as to meet the demands of a

knowledge society. E-governance can be viewed as providing citizens with the ability to choose the manner in which they interact with governments, and ensuring ICT can be used effectively to improve the flow of information and citizen to government relationships.

Digital or Electronic Government (E-government) is the use of ICT to promote and motivate a more efficient and cost-effective government; facilitate more convenient government services; allow greater public access to information; and make government more accountable to their citizens. Digital government comprises electronic service delivery, electronic democracy, and e-governance (digital support for policy making and the policy process).

It should be noted that digital government (or governance) is not primarily a technical exercise, but rather an attempt to improve the political and social environment through the utilization of ICT. Introduction of automation to the public sector will not automatically create a better or more open government - unless it is based on policies to promote the effective utilization of technology. Digital government inevitably needs to take into consideration issues such as new models of policy formulation; alternative forms of citizenship; different patterns and trends of relationship and power; new solutions for economic development; and alternative approaches for connecting people to the political process. The adoption of digital government poses fundamental questions - such as:

- the ways in which governments can act effectively in this new environment - in the interests of their citizens
- the role and capacity of government
- the relationship they have with their citizens and with other parties
- the boundaries of community and therefore of representation and citizenship that go to the heart of democratic theory

Extensive research has been conducted (and is still being carried out) by various practitioners in an attempt to answer some of these questions (e.g. [6], [15], [5], and [22]). Furthermore, certain advisory and interest groups have been formed in order to provide answers to some of these questions (e.g. International Centre for e-Governance – www.icegov.org).

It is strongly believed that the essence of governance/government centres on relationships. Hence an effective model for developing e-governance or e-government strategies needs to consider the connectivity between different views and domains of government - one that views e-Government as a complex system (or a hub) whereby ICT solutions are introduced so as to interconnect the various domains of governance, in order to facilitate increased operational efficiency; enhanced economic development; improved service delivery; redefined communities; enhanced citizen participation;

improved policy formulation; and global interconnectivity

In general, digital government can include the following practices:

- automation of government systems and the online delivery of services
- widespread adoption of network-based technologies and the migration of government to the Internet environment
- the application of electronic capabilities and practices to government to reduce costs, reduce fraud, and increase efficiency
- the use of ICT to foster economic growth and conduct business
- improvement or re-engineering of the structures of government and the nature of public administration
- use of ICT to foster democracy and citizen engagement and improve political accountability

As the application of ICT in governments within the developing nations including the Asia Pacific region becomes more widespread (see [5], [6]), we begin to observe a progression through the various stages of electronic government – which includes:

- improving internal functional efficiency through the application of ICT
- improving internal communications (through the application of electronic mail) and introducing workflow management systems for increased process efficiency
- putting in place applications that would not only enable citizen participation through feedback, but would also allow for transactions between citizens to government (C2G), businesses to government (B2G) and government to government (G2G).
- introducing digital democracy - technological solutions that enable participatory action and democratic processes
- introducing integrated electronic or digital governance

3. The Potential Impact of Digital Government

3.1 Models for Analyzing the Impact of Digital Government

A study of various debates over the implications of electronic or digital government (e.g. 6, 2000, Asgarkhani 2002) indicates that there are at least four schools of thought (models) which include:

- pure optimism
- optimism with some concerns
- pessimism
- technology viewed as a tool only - but not a

driving factor on its own

The optimists argue uncompromisingly that the use of technology in governance represents a major once-and-for-all improvement in the capabilities of governance through a more effective management of all domains [17]. The only cost is considered to be the investment and the day-to-day operational running costs. It is believed that these systems can reduce the costs of decision making, management and day-to-day operational activities (such as acquiring, ordering, coding, organising, selecting, managing and using information) steadily over time. That is to say, the initial investment costs would be compensated through the cost savings and efficiency gains that are likely to be achieved over the lifetime of the systems [14]. This optimistic view appears to be based on the classical cybernetic theory [22] – a theory that views information as control. It argues that information decreases uncertainty; slows entropy; and increases system control by enabling feedback and deviation correction - and that more information enables more control [25].

The second group (optimists who have some concerns) accept at least the possibility of greater control, quality and rationality in decision-making. However, they argue the efficiency gains achieved through digital government come at a price. More specifically, they believe unless safeguards are put in place, digital government systems may result in compromising citizens' rights such as:

- the right to individual liberty and privacy
- the right to influence governmental decision-making [25], [13]
- losing control over politicians' decision-making agendas [24]

The pessimists argue that e-digital government will actually compromise the quality of decision-making. They are concerned that excessive demand for policy analysis based on many categories of information will cause delays in action – “paralysis by analysis.” There is the fear that due to mechanical rule following, as suggested by overly simple data interpretations, overly simple modelling, and by overly simple expert system flows from analysis to recommendation, the cultivation and the exercise of judgement in decision-making will be downplayed. As you have probably guessed, this view rejects the cybernetic theory - that information is control.

The last group view technology as a tool and argue that the impact of ICT solutions cannot be viewed in isolation where it concerns technical or political rationality of decision-making. They view both continuities and changes in governance as being driven socially and politically, not by technology itself. Technology is seen as a tool for either changing or preserving the *style of governance* – e.g. *conservative* and *radical* styles of governance [10], [7].

Each theory that has been mentioned above has some empirical support - although most empirical studies

have been of a rather limited scope and are not in general designed to test, let alone falsify these rival theories [25]. The study of these models raises many questions, such as:

- how are these schools of thought to be appraised?
- is it possible to favour one that is correct or at least not yet falsified?
- is it perhaps possible to allocate them to different domains?
- is there more than one valid view?

It is fair to say that providing an answer to these questions in a unified manner that applies to every situation across the board would be unrealistic. The answer could depend on numerous factors such as social and cultural aspects; the technological infrastructure; past experience with the application of ICT; the level of education and interest in the political process and so forth.

3.2 Social, Cultural and Ethical Challenges of Digital Government

Some of the perceived social implications of digital government can include:

Information Security - technological advancements allow government agencies to collect, store and make available to others online data on individuals and organizations. Furthermore, citizens and businesses expect to be allowed to access data in a flexible manner (access to data anytime, from any location). Meeting these expectations comes at a price to government agencies in regards to managing information – including: ease of access; data integrity and accuracy; capacity planning to ensure the timely delivery of data to remote (possibly mobile) sites; and managing the security of corporate information [3].

Impact on Jobs and Workplaces - in the early days of computers, management scientists anticipated that computers would replace human decision-makers. However, despite significant technological advances, this prediction is no longer appearing to be a mainstream concern. At the current time, one of the concerns associated with computer usage in any organization (including governments) is the health risk – such as injuries related to working continuously on a computer keyboard). Government agencies are expected to work with regulatory groups in order to avoid these problems.

Impacts on Individuals' Rights and Privacy – as more and more companies and government agencies use technology to collect, store, and make accessible data on individuals, privacy concerns have grown. Some of these concerns are related to maintaining the individual privacy of employees as well as citizens. Some companies choose to monitor their employees' computer usage patterns in order to assess individual or workgroup performance [4]. Technological advancements are also

making it much easier for businesses, government and other individuals to obtain a great deal of information about an individual without the individual's personal knowledge. There is a growing concern that access to a wide range of information can be dangerous within politically corrupt government agencies.

Potential Impacts on Society – despite some economic benefits of ICT to individuals, there is evidence that the computer literacy and access gap between the haves and have-nots may be increasing. Education and information access are more than ever the keys to economic prosperity, yet access by individuals in different countries is not equal - this social inequity has become known as the digital divide [1].

Impact on Social Interaction – advancements in ICT and web-based technology solutions have enabled many government functions to be automated and information to be made available online. This is a concern - in particular, considering those cultures that place a high value on social interaction.

4. The Digital Divide and e-Readiness

As discussed in Section 1, change is taking place on a global level, and the pace of change is accelerating. Changes include the rapid expansion of the Internet and the network economy; increasing business and individual use of wireless communications; access devices and new ICTs; the convergence of technologies, services and markets; globalisation of markets and trade; greater competition; mergers and acquisitions and their impacts on the restructuring of industry and markets. These changes are transforming all areas of society, work, business, and government [23]. Today, access to information and communication technologies (ICTs) is critical for economic and social development. Network economics means that the more that ICTs are adopted, the greater the value to all users.

However, there is some concern as there appears to be conflicting scenarios for the global information society. There is much optimism that we are facing a myriad of digital opportunities where the means exist to broaden participation in the network-based economy and to share its benefits. At the same time, differences in diffusion and use of ICTs and electronic networks appear to be deepening and intensifying the socio-economic divisions among people, businesses and nations. The digital divide takes various forms, and impacts different individuals, businesses, regions and nations differently - including:

- divides between countries
- social divides within countries
- divides within countries related to income, education, age, family type, location and so on
- business divides related to sector, region, firm size and so forth

The digital divide can limit the success of digital government solutions. Even though governments in some developing countries receive funding and support for introducing digital government solutions, the effectiveness of these solutions are limited – unless the barriers to e-readiness within the nation are addressed.

Typical causes of digital divide that can also limit successful implementation of digital government solutions can include:

- lack of telecommunications and network infrastructure
- limited PC access
- lack of financial resources for developing infrastructure
- lack of ICT literacy
- limited Internet access
- cultural resistance
- high access costs to global networks and the Internet
- high cost of business investment
- strategic business impediments - applicability, the need to reorganise, the need for skills, security and privacy considerations

A review of some of the studies on the digital divide and e-readiness (e.g. [19], [1], [8], [12] and [20]), indicates that there are significant differences with regards to the state of the application, network economy and digital government worldwide. Let us look at a small sample of these studies.

In 2000, the META Group [12] examined the digital commerce competitiveness of 47 countries in an attempt to establish a digital economy index. The author of this study, Howard Robin wrote, “Traditional industrial-age measures of production and performance have lost relevance in the information age. Currently, information processing capability is a better indicator of national competitive advantage.” The research by the META Group ranked 47 countries in five different categories in order to establish an overall ‘information age technological competitiveness.’ These categories included knowledge jobs; globalisation; economic dynamism and competition; transformation to digital economy; and technological innovation capacity. Table 1 outlines some of the results as they concern some of the countries within the Asia Pacific region (including New Zealand and Australia).

Next, we look at the 2002 Information Society Index (ISI). This project considered 23 parameters so as to compile a ranking list of 55 countries (that accounted for 98 percent of the ICT solutions within 150 countries). The countries that were featured in the ISI index were classified under four categories. These four categories (along with examples of Asia Pacific countries that featured in the 2002 ISI index) are as follows:

Skaters – these countries are in a strong position to take full advantage of the information revolution, as they appear to have advanced ICT and social infrastructures.

Asia Pacific countries that were classified under this category included: Australia (ranked 9th); Taiwan (ranked 10th); Hong Kong (ranked 11th); Japan (ranked 12th); and Singapore (ranked 13th).

Striders – are countries that appear to be moving purposefully into the information age, with much of the necessary infrastructure in place. This category included New Zealand (ranked 17th) and Korea (ranked 18th).

Sprinters – are countries that are moving forward in spurts before needing to catch their breath and shift priorities due to economic, social and political pressures. Malaysia (ranked 30th) was the only Asian country in this group

Strollers – are those moving ahead but inconsistently, due to limited financial resources in relation to their vast populations. Table 2 displays the list of Asia Pacific countries featured in the ISI. Countries that were considered under this category were: Philippines (ranked 45th); Thailand (ranked 46th); China (ranked 52nd); India (ranked 53rd); Indonesia (ranked 54th); and Pakistan (ranked 55th).

The first eight countries in the 2002 ISI index were: Sweden; Norway; Switzerland; the United States; Denmark; the Netherlands; the United Kingdom; and Finland.

Country	Ranking (out of 47)					
	Overall Index - technological competitiveness	Knowledge Jobs	Globalisation	Economic Dynamism and Competition	Transformation to Digital Economy	Technological Innovation Capacity
Japan	2	38	5	30	13	1
Australia	8	2	19	20	6	14
Taiwan	10	26	23	3	10	8
New Zealand	11	9	20	16	7	37
Hong Kong (SAR)	15	30	6	8	27	43
Singapore	17	25	14	6	19	33
Philippines	25	1	35	34	32	38
Malaysia	33	33	37	27	14	42
India	34	8	41	43	42	17
China	37	46	34	36	12	5
Korea	38	39	43	44	23	9
Thailand	46	43	36	40	47	39
Indonesia	47	47	45	47	45	11

Table 1. Analysis of META Group Research by Technological Competitiveness

A survey of online governance conducted by UNESCO [20] outlines a number of other key statistics with regards to e-governance. Some of the findings of this research (with a focus on the Asia Pacific region) have been summarized below.

- About 84% of the respondents (Asia Pacific) had a government website (7% did not reply, and 7% did not have a website)
- 67% of the respondents provided free government information online whilst 20% charged fees for some information
- With regards to ICT access, 38% of the respondents provided public kiosks for online access; 22% subsidized the Internet access; and 26% subsidized the purchase of computers
- Under 40% of the respondents used smart card technology
- The key inhibitors to the development of digital governance in developing or underdeveloped nations included: lack of infrastructure (60%); lack of resources (47%); low level of ICT literacy (33%); lack of awareness at policy level (20%); low public incentives (27%); and low internet penetration (20%).

Overall, a relatively small sample of research outcomes (concerning e-government) cannot be applied to all countries. However, it appears that many countries (including some Asia Pacific governments) are at the initial or half-way stages of adopting ICT solutions in order to introduce digital government.

As mentioned earlier, unless the barriers to e-readiness within the nation are addressed, the success of digital government would be limited.

5. Citizens' Perception of Effectiveness

In December 2001 the Christchurch City Council introduced an Electronic/Digital Governance initiative called the "eCouncil Project." Their initiative reflects trends in local government organizations worldwide in the development of ICT based solutions in order to enhance communication and information flow between government and citizens. The Council aims to utilize ICT through the implementation of its Digital Governance initiative to facilitate improved two-way exchange of information and enhance its public image as a professional customer service oriented organization. An analysis of the eCouncil project shows the Council to be closely following recent worldwide trends, and shows significant progress has been made towards the implementation of electronic governance. They acknowledge that successful implementation of electronic governance does not result in merely automating the collection and distribution of information, but results in the flow of useful information between the government organization and its citizens.

The Council itself measures the success of the eCouncil Project on an ongoing basis – by looking at: website hits; customer feedback; and quantifiable efficiency benefits.

- **Website Hits:** These are monitored to determine the utilization of services. In May 2003 the Council's main service website scored 350,000 hits. The 30 other websites maintained by the Council scored between 150,000 and 180,000 hits.
- **Customer Feedback:** The Council pays particular to customer feedback from facilities on the service websites. Feedback is assessed to measure customer satisfaction and service level impact. Suggestions may also result in changes to the services provided.
- **Quantifiable Efficiency Benefit:** Services provided by the eCouncil project are intended to contribute to a reduction in operating costs. Services must continue to meet the desired levels of efficiency – which includes cost savings, time saving, and service level impact.

Interviews were conducted to assess public knowledge and opinion of the Christchurch City Council's online services website. Participants were chosen from various age groups -more specifically, 64% of the respondents were 18-34 years, 26% were aged between 35 and 49 and the rest were over 50 years.

It appeared that 84% of the respondents were aware of the digital governance services that are provided by the Council online. All participants viewed the electronic delivery of services as being useful. They emphasised ease of access and flexibility (among others) as being their key reasons.

Results indicated that the majority of respondents (39%) were aware of the Council's website through word of mouth. Web surfing or search engines were second (30%) whilst 15% of respondents knew of electronic service delivery through advertising. However, 16% did not know at all that the Council provides online services.

Those respondents, who provided additional information, seemed to have used the following services:

- Bus Timetables
- Job Listings
- Population Statistics
- Street Maps
- Rates Information

All respondents (who had used the website services) rated the Council's online service as being effective. Those who provided additional information stated the following reasons for its effectiveness:

- easy to navigate
- relevant information
- information could successfully be retrieved
- fast page loading
- high level of usability
- immediate access to required information
- time saving

One respondent stated the website did not work correctly. Other respondents stated other difficulties – such as:

- could not find the required information
- download and access time was slow
- navigation difficulties

These results indicate that overall, users of the website are satisfied with the information provided and the site's level of usability.

On the issue of the Council services that would be suitable for online delivery, respondents suggested the following services:

- rates payments
- other council fee payments
- additional contact information for Council service departments
- application forms for services
- rubbish collection information
- interactive services: online forums and discussion groups
- multimedia: streaming video and audio of local events
- online voting facilities

Participants were asked to state their concerns about using online services. Results show data security is the greatest concern for customers (43%). Concerns about confidentiality of data came second (17%). 6% of the respondents had concerns about document compatibility and another 6% were unhappy about the speed of access. However, 28% of participants had no concern about using Council's online services.

On a scale of 1 to 10, (10 being the greatest), the rating for website content was 8.2.

Overall, on the scale of 1 to 10 (10 being highly desirable), results showed the average rating of the Council's online services was also 8.2. Furthermore, respondents gave the importance of access to online services (in general, not just the Council's services) a score of 7.6 (10 being highly desirable).

Even though these results of this pilot study are not to be considered as final, it appears that Council's digital government solutions are rated as being highly effective and useful. This is consistent with New Zealand's state of adopting ICT solutions and e-readiness - as outlined in Section 4. However, these results can not be generalized over other digital government solutions provided in other countries.

6. Conclusions

Within the past few years, much has been argued about the use of ICT and its effectiveness in the process of governance.

We examined the effectiveness (i.e. success and usefulness) of digital government by considering different aspects of effectiveness – including:

- *The view of management and ICT strategists with regards to the implications of digital government* – ranging from the optimists who view digital government as being an effective tool without any concerns to those who view technology as a tool only (arguing that technology on its own cannot be a driving force for effectiveness).
- *The implications of digital divide and e-readiness* – the effectiveness and success of digital government in a country rely on its state of e-readiness and the ways in which the barriers to the 'digital divide' can be overcome.
- *The customers'/citizens' view of the usefulness and success of digital government* – a pilot study of digital government at the Christchurch City Council (CCC) indicates that local citizens rate digital government solutions (offered by the CCC) as being effective. However, these results are not final and cannot be generalized over other digital government solutions that are put in place in other countries.

Overall, technical innovation on its own is not enough to drive the process of developing effective digital government solutions. More specifically, access to the right technology for delivering digital government is essential but insufficient. Even though most of the shortcomings (as they concern the effectiveness of digital government) can be resolved by improving the technology infrastructure, technology by itself does not necessarily result in better, more efficient governance in governments. Technological advancements are only effective if they are considered alongside other key parameters such as social structure; cultural values and attitudes; governance process re-engineering within governments; and ethical issues.

There are a number of key questions that can be considered in order to facilitate the development of more effective digital government solutions [18] including:

- What are our reasons for pursuing e-Government? Are we aware of the challenges of the path to e-Government and our infrastructure needs?
- What is our clear vision for e-Government?
- What are the priorities that we have considered with regards to e-Government services?
- What is the type of e-Government that we are ready for? This involves reviewing current connectivity, telecommunication infrastructure, the political will for e-Governance, information policy and so forth.
- Do we have a methodology for selecting, planning and managing e-Government projects?
- Did we consider a thorough plan for managing change?
- What are the tools and metrics that we have thought of in order to be able to measure

progress/success?

- What would be a model (methodology) for managing relationships with the private sector?
- How would our e-governance model improve citizen participation in public affairs?
- What are our priorities with regards to global connectivity to other governments?

Acknowledgements

My thanks to *Rochelle Jones, Andrew Scott, Alex Lissaman, Terry Moon, Fiona Rice and Saisri Pulkanam* (my research assistants) for their contribution towards the study of digital local government services in New Zealand.

References

- [1] Accenture, (2001a) "e-Government Leadership – Realizing the Vision", [online] <http://www.digitalopportunity.org>.
- [2] Accenture (2001b) "Governments Closing Gap Between Political Rhetoric and e-Government Reality", [online] http://www.accenture.com/xd/xd.asp?it=enWeb&xd=industries/government/gove_study.xml.
- [3] Asgarkhani, M. (2001) "Managing Information Security: A Holistic Business Process", *Proceedings of the 20th IT Conference*, Sri Lanka, July, 2001, pp93-100.
- [4] Asgarkhani, M. (2002a) "Strategic Management of Information systems and Technology in an e-World", *Proceedings of the 21st IT Conference*, Sri Lanka, pp103-111.
- [5] Asgarkhani, M. (2002b) "e-Governance in Asia Pacific", *Proceedings of the International Conference on Governance in Asia*, Hong Kong.
- [6] Asgarkhani, M. (2003) "A Strategic Framework for Electronic Government" *Proceedings of the 22nd National IT Conference*, Sri Lanka, pp57-65.
- [7] Bijker, W.E. (1997) "Of bicycles, bakelites and bulbs: toward a theory of sociotechnical change", *Inside Technology*, Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1997.
- [8] COMNET-IT (2002) "Country Profiles of E-Governance", [online] <http://www.digitalopportunity.org>.
- [9] Heeks, R. (1999) "Reinventing government in the information age: international practice in IT-enabled public sector reform", Routledge. London.
- [10] Mackenzie, D., and Wajcman, J. (1985) "The social shaping of technology: how the refrigerator got its hum", Open University Press, Buckingham.
- [11] Massetti, B. (1998) "An empirical examination of the value of creativity support systems on idea generation", *Management information systems quarterly*, 20, 1998-1, pp83-98.
- [12] META Group (2000) "The Global E-Economy Index", [online] <http://www.ecommercetimes.com>.
- [13] Raab, C. (1997) "Privacy, information and democracy", in Loader BD, ed, 1997, *The governance of cyberspace: politics, technology and global restructuring*, Routledge, London, pp155-174.
- [14] Reschenthaler, G.B., and Thompson, F. (1996), "The information revolution and the new public management", *Journal of public administration research and theory*, Vol 6, No. 1, pp125-143.
- [15] Samaranyake, V. K. (2003) "The Reality of Digital Government", *Proceedings of the 22nd National IT Conference*, Sri Lanka, pp1-9.
- [16] Stevens, J.M. and McGowan, R.P. (1985) "Information systems for public management", Praeger, New York.
- [17] Tapscott, D. (1997) "The digital media and the reinvention of government", *Canadian public administration*, 40, pp328-345.
- [18] The Working Group on E-government in Developing World – WGEDW (2002) "Roadmap for E-government in Developing World", [online] <http://www.digitalopportunity.org>.
- [19] UN E-Government Report (2001) "Benchmarking E-Government: A Global Perspective-Assessing the UN member states", UN Publication, [online] <http://www.upan1.org/egovernment2.asp>.
- [20] UNESCO/COMNET-IT (2000) "Global Survey of Online Governance", [online] <http://www.digitalopportunity.org>.
- [21] Wescott, C.G. (2001), "E-Government in the Asia-Pacific Region", UNPAN Asia Pacific, [online] <http://www.unpan.org>.
- [22] Wiener, N. (1984) "Cybernetics: the emerging science at the edge of order and chaos", Simon and Schuster, New York.
- [23] Workshop – Digital Divide – OECD (2000), "The Digital Divide: Enhancing Access to ICTs", [online] <http://www.oecd.org/dataoecd/22/11/2428340.pdf>.
- [24] Zuurmond, A. (1988) "From bureaucracy to infocracy: are democratic institutions lagging behind?", *Public administration in an information age: a handbook*, IOS Press, Amsterdam, pp259-272.
- [25] 6, Perri (2000) "E-governance: Weber's Revenge?" *Proceedings of the Annual Conference of the Political Studies Association*.



Non-intentional Cooperative Behaviour for an Agent Based Intelligent Environment

R.A. Chaminda Ranasinghe
University of Colombo, School of Computing,
No. 35, Reid Avenue,
Colombo 7, Sri Lanka.

chaminda@interblocks.com

Ajith P. Madurapperuma
Department of Computational Mathematics,
Faculty of IT, No 100/A, D S Senanayake Mawatha,
Colombo 8, Sri Lanka.

ajith@itfac.mrt.ac.lk

Abstract

Our work with the AAANTS (Adaptive, Autonomous, Agent colony interactions with Network Transparent Services) project applies knowledge gathered from the study of natural community life styles, for e.g. Ants, to developing methodologies for intelligent behaviour in agent-based synthetic ecosystems. This paper discusses the feasibility and implications of using non-intentional interactions among entities in a multi-agent system to coordinate collective behaviour as opposed to agent interaction techniques adapted in common deliberative multi-agent systems. The implementation of AAANTS model within the framework of an Intelligent Environment has confirmed the ability of multitude of loosely coupled egalitarian collection of agents to depict adaptive cooperative behaviour using a non-intentional communication model.

Keywords: Multi-agent systems, Emergent behaviour, Intelligent Adaptive Systems, Distributed agent architectures, Ubiquitous computing, and synthetic ecosystems

1. Introduction

Ants like many other insect species, occupy a central place in artificial life due to their individual simplicity combined with their relatively complex group behaviour [12]. Ant colonies have evolved means of performing collective tasks, which are far beyond the capacities of their constituent components. They do so without being hard-wired together in any specific architectural pattern, without central control.

According to [7], the amazing success of the Ants is due to the swiftly applied and overwhelming power arising from the cooperation of colony members. Ants, like humans, succeed because they talk so well [7]. Further, an Ant colony can be regarded as a super organism where it can be analysed as a coherent unit and compared with the organism in design of

experiments, with individuals treated as the rough analogues of cells [7].

AAANTS model conceptualises a multi-agent system, i.e. a *Colony of Agents*, consisting of autonomous software components resembling agent characteristics that work in harmony and synergy to achieve community wide goals [13]. At present, an introductory level definition can be given to an agent as an entity with perceptions, goals, cognition, actions, and domain knowledge, situated in an environment [17]. AAANTS model uses the community life style of insects as a metaphor with further inspiration from “The Society of Mind” theory [10]. The components in the AAANTS model can be broadly segmented into a collection of agents and a distributed collection of embedded services. The services act as a neural extension to the agents in providing real-time sensory information from the natural environment.

We have positioned AAANTS as a hybrid model due to the presence of features from both deliberative and reactive paradigms. The hybrid nature of the system is prominently demonstrated in areas of knowledge representation, agent interaction, and adaptive nature in terms of learning and periodic evolution.

The AAANTS model has gone through several stages of modelling, framework development, knowledge-representation techniques, learning methods and cooperation strategies, during the past. We have found out that the coordination strategy remains as the core focus during any team work environment and other methodologies should be moulded to complement it. In the AAANTS model, we try to make the interaction simple by eliminating explicit active communication adapted by most deliberative agent models [14]. In the rest of the paper, we describe the nature of communication found in the AAANTS model and how it contributes to the overall cohesiveness and synergy expected by a multi-agent system.

2. Reasons for Cooperation?

Cooperative behaviour among collection of individuals has been the cornerstone for the success of human beings' ability to conquer complexity. This is evident when analysing many important historical moments ranging from wars to innovative designs. A multi-agent system too, is composed of several units of autonomous entities that interact to achieve a collective goal. Without cooperation, an agent is merely an isolated individual, closed into its perception-deliberation-action loop [4]. Therefore, co-operation is an important factor to the success in a multi-agent system.

Further, we need to clarify the ambiguity of the terms co-operation and co-ordination. Co-ordination is a process which agents engage in order to ensure a community of individual agents act in a coherent manner. Co-ordination, in turn, may require co-operation; but it is important to emphasize that co-operation among a set of agents would not necessarily result in co-ordination; indeed, it may result in incoherent behaviour [11]. According to [11], the reasons for co-ordination are, preventing anarchy or chaos, meeting global constraints, distributed expertise, resources or information, dependencies among agent actions, and efficiency.

3. Mechanisms for Cooperative Behaviour in Multi-agent Systems

Communication facilitates sharing of intelligence, negotiations, collaboration and co-ordination. Software agents use a communication language for similar purposes. The main reason for communication may vary depending on the purpose of an agent's existence. The main substance of agent communication is defined in an Agent Communication Language (ACL) [8]. An ACL enables software agents with ontological [3] [18] similarities to communicate with each other via an extensible set of "performatives" expressing beliefs and attitudes towards some information elements. A performative specifies the format of any given message and dictates how an agent should respond to messages. Two popular communication languages are the Knowledge Query and Mark-up Language (KQML) and Agent Communication Language (FIPA ACL) [8] [9] [2].

The break down of predefined tasks found in cognitive agents can be managed by centralising the allocation process or by distributing it among all the agents concerned. The centralised and distributed approaches are concerned with the allocation of tasks by cognitive agents capable of intentionally communicating with each other [4]. In contrast,

reactive agents use the concept of signals, which are non-intentional forms of communication, sent by diffusion and propagation into the environment. The proposed AAANTS model conceptualises its communication model based on these elementary form of communication found in reactive paradigms.

According to [6] [16], Agent Communication Languages can best be thought of as consisting of three parts - its vocabulary, an "inner language" such as KIF (Knowledge Interchange Format), and an "outer" language such as KQML or FIPA-ACL. For example an ACL message can be a KQML expression in which the "arguments" are terms or sentences in KIF formed from words in the ACL vocabulary.

According to [1], KQML and FIPA-ACL use may be too complicated for some kinds of applications that do not need speech acts and logic to carry out their negotiations. We embrace this observation for the proposed methodology of interaction in the AAANTS model. The AAANTS model possesses capabilities to simplify the coordinated interaction by eliminating explicit active communication adapted by most deliberative agent models.

Multi-agent systems based on the reactive paradigm do not make use of ontological basis of knowledge sharing to the extent used by cognitive / deliberative paradigms. The proposed AAANTS model though being simple in terms of depth, flavour and nature of inter-agent communication, uses ontologies during communication.

4. Non-Intentional Cooperative Behaviour

The agents in the AAANTS Colony are segmented into groups that share common behaviour and ontologies. In other words, a group of agents is responsible for a spectrum of homogeneous behaviour. For example, the behaviour of activating a lamp is undertaken by a collection of agents that may be triggered by different environment conditions such as darkness, during user trying to read a book or intruder detection. Therefore, the state of the environment sensed through the embedded services is responsible for activating suitable agent behaviour.

4.1 Message Structure

According to [2], a typical agent communication language can be divided into three layers consisting of Content Layer, Message Layer and Communication Layer (*Figure 1*). In the AAANTS model too, we have used similar segmentation to handle complexity during agent interoperability. Initially a communication layer is used for interaction among components such as agents, services and administrator/monitoring tools. This layer combines distributed locations within the

agent colony and services embedded in the environment.

The Message Layer consists of encapsulated message packets that contain a header and content information. Sensory signals published by distributed services and actuator signals published by the agents are disseminated in the network on a predefined subject. The subject together with other meta information is represented in the header portion of the message. The agents and services could publish, subscribe and intercept messages on a subject of interest. This concept adheres to the Observer Pattern as described in [5]. A subject simply represents a homogeneous collection of sensations or behaviours that is attached to each message. Subjects are organised in a hierarchy so that a consumer listening to a parent subject may intercept all inherited messages classified under the parent and can be formulated as listed below.

$S1 = \{x: x \text{ is a subject}\}$

$S2 = \{x: x \text{ is a subject}\}$

$S2 \subseteq S1 \Leftrightarrow (\forall x, x \in S2 \Rightarrow x \in S1)$

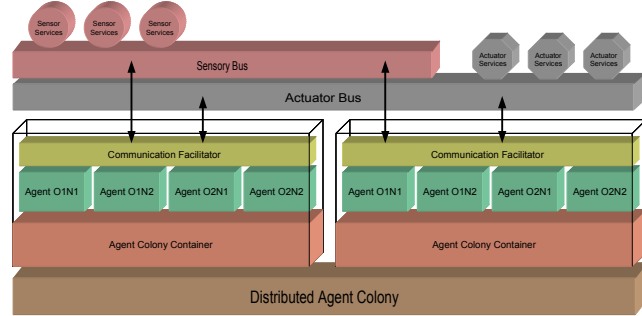


Figure 1: Agent interaction using sensory and actuator messages.

Published messages are not naturally retained within the network for later consultation. Therefore, the AAANTS framework has provided a service called Message Queue Server (MQS) to retain the history of published messages. This essentially acts as a repository of all sensations and actuator messages that have taken place within a specific period in time. Agents can communicate with the MQS to query recent patterns of data. Each agent undergoes an adaptation stage with the intention of improving their behaviour to evolving environment conditions. The information stored in the MQS performs an analogous function to that of pheromones used in insect colonies. Pheromones are chemicals deposited by individual insects in order to exchange information among individuals and are evaporated temporally. Similarly, the sensory and actuator information captured by MQS are dissipated temporally.

4.2 Knowledge Representation

Agents as discussed earlier, are autonomous entities that respond to environment sensations while maintaining coherent knowledge structures relevant for its behaviour. Agents in the AAANTS model perform the inference related activities individually by matching information gathered from the surrounding with the frame-based knowledge structures in possession [15]. Further, these frame-based knowledge structures are modified using Reinforcement Learning techniques based on varying environment state that reinforce a certain behaviour. Actuator channels too can be used as input to agents since behaviour of some agents can act as sensations to others. For example, activating the behaviour of “opening of a door” can act as a sensation to trigger activity on other services such as lighting, air conditioners, electric appliances etc.

Agent behaviour naturally does not solely depend on another for activation, since other environment conditions need to be consulted. For example, a group of agents may have adapted to a relationship of a human entering a room in summer with that of activating the air conditioner. Such basic behaviour makes agents naively adapt repetitive patterns without considering other complementary factors in the environment. A better solution would be to gather other complementary variables from the environment and adapt to changing situations in a real-time manner. With reference to the above example, it would be more appropriate to activate the air conditioner taking into consideration the environment variables such as temperature, humidity, time, and other predictive behaviour. In addition, the user in weekends might spend only few minutes in the specified environment for some mundane activities and might not need the air conditioner to be turned on.

4.3 Reason for Non-Intentional Behaviour

We describe the cooperative behaviour of agents in the AAANTS model as “**non-intentional**”, since there does not exist any intentional direct communication among agents using an accepted agent communication language. Agent interaction is facilitated by message exchanges disseminated in the network where the interested agents are responsible for intercepting and processing the published messages to exhibit further behaviour. This methodology enables information sharing among a group and the ability to influence behaviour on others without explicit knowledge about the participants: thereby making the interactions, non-intentional. We find this methodology having

resemblance to the interaction mechanisms found in insect colonies with the use of chemicals such as pheromones.

5. AAANTS Coordination Model

As discussed earlier, we have defined the AAANTS system as a multi-agent system where a community of agents achieves goals collectively. So the agents will time to time submit individually decided actions to overcome needs of the community. When several urgent needs occur at once, there must be a way to resolve conflicts. One scheme for this might use some sort of central market place, in which the urgencies of different goals compete and the highest bidder takes control [10]. This strategy may fail since extent of achievement in the selected goal may not be defined. Another way is to use an arrangement called cross-exclusion, which appears in many portions of the brain [10]. In such a system, each member of a group of agents is wired to send “inhibitory” signals to all other agents of that group which makes them competitors. When any agent of such a group is aroused, its signals tend to inhibit others. This leads to an avalanche effect, as each competitor grows weaker; its ability to inhibit its challengers also weakens. The result is that even if the initial difference between competitors is small, the most active agent will quickly lock out all the others.

So cross-exclusion is one of the methods that can be used to regulate levels of activities in an agent society. But cross-exclusion can make some selected goal to totally dominate the agent functionality through inhibition.

The AAANTS model is composed of a collection of agents, each responsible for a defined type of activity. For example, with reference to the (Figure 2), the movement of a Robot with four wheels and two motors on either side is controlled by four basic behaviours such as forward, turn left, turn right and stop. These four movements could be sequenced in various permutations to depict wide range of synchronised and intelligent activities.

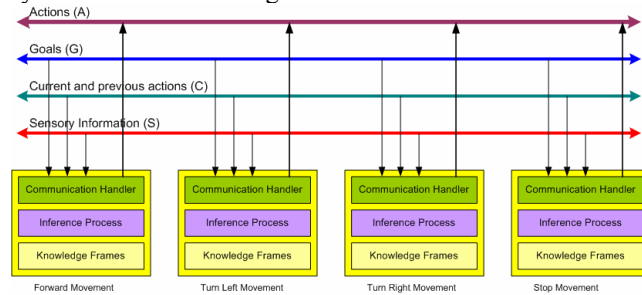


Figure 2: AAANTS Coordination Mechanism

We have discussed the AAANTS communication model as non-intentional due to absence of direct intended communication among agents. Therefore, the synchronisation related information should be kept at each agent that participates in an emergent behaviour. This information is stored in frame based knowledge structures located at each agent. The information related to sensations, goals, current and previous actions are continually published through the communication channel. When an agent intercepts these signals through the communication handler, these are matched against the knowledge structures by the inference process. The inference process should select the most appropriate behaviour.

The inference process would take into account the current goal, on-going and previous activities and environment sensations. The same behaviour for example “Move Forward” can be depicted under different goals such as object tracking, move an object from source to destination or in order to reach the power source for recharging.

6. Implementation of the Non-intentional Model of Communication

The implementation of the AAANTS model initially has focussed on developing a framework to facilitate the design objectives. The framework mainly focuses on providing facilities to a multi-agent system such as process management, communication, agent life-cycle management, persistence management, mobility and security to a multi-agent system [14]. Implementation details of communication sub-system are of main concern within this paper.

6.1 Communication Layer

As described earlier, the AAANTS communication subsystem can be segmented into Content, Message and Communication Layers together with some system level services. The Communication Layer is implemented using UDP multicasting. Multicasting enables information publishers to disseminate a single message to multiple subscribers thereby eliminating redundant, retransmission of messages found in a unicast protocol such as TCP/IP. We have used two multicasting groups to separate messages into sensory and actuator origin as depicted in (figure 1). This separation has been intentionally performed due to the high traffic rate in the sensory channel.

6.2 Message Layer

The Message Layer is placed on top on the Communication Layer to provide proper encapsulation

of sensory and actuator signals as messages. The messages are created by the publishers and intercepted by the subscribers. Each message contains a header and a data portion. The header contains information such as subject, originator ID, verification data and sequence number. The main publishers of the sensory channel are the heterogeneous sensory services that capture environmental sensations such as real-time video, audio, voice recognition, temperature and motion. The adjective “heterogeneous” is intentional in generalising the services because of variety of sensations, platforms and application programming languages (C, C++ and Java). The primary consumers of the sensory channel are the egalitarian collection of agents that relentlessly listen for messages published by the sensory services. The next channel as depicted in *figure 1* is actuator channel, which mainly carries actuator signals published by the agents. The agents perform real-time processing of signals against their knowledge bases to publish the inference as messages that can activate behaviour in processes embedded in the environment, called actuator services. The agents that belong to a homogeneous group recursively become listeners to the messages in the actuator signal channel. This enables agents to give real-time sequence of inter-dependent activities that has been learned in the past.

6.3 Content Layer

The Content Layer is embedded with the Message Layer and mainly focuses in the data portion of the message. The content is based on XML that has the natural advantage of describing various types of content. Both agents and services possess XML parsers to create and extract information from the content layer. We have included further functionalities in the XML parsers used by agents to handle disparate and unpredictable patterns of content.

7. The Prototype

We have initially created several wrappers in C++ and Java to represent the Communication Layer that handle multicast messaging within a distributed network. These wrappers adhere to both Proxy and Observer Pattern as described in [5]. A message wrapper is used by another library that offers a façade to construct and extract messages that represents the message layer, which is further extended to handle XML content manipulation. These libraries were further amalgamated with other processes to build up the AAANTS framework.

We have developed an initial simulation (*Figure 3*) to represent an Intelligent Room, using several

sensory and actuator services. We have selected some basic level sensory services such as motion sensor, sound sensor, vision camera for gesture recognition, infrared sensor for remote control commands and a voice recognition engine developed on IBM ViaVoice engine. Each of these services, though heterogeneous in functionality, merges into a single level of interaction because of XML based content layer. We have also developed some actuator services based on a robotic toolkit (Lego Mindstorms), voice synthesis engine (based on IBM ViaVoice), text message sender and X10 based electronic appliance controller.

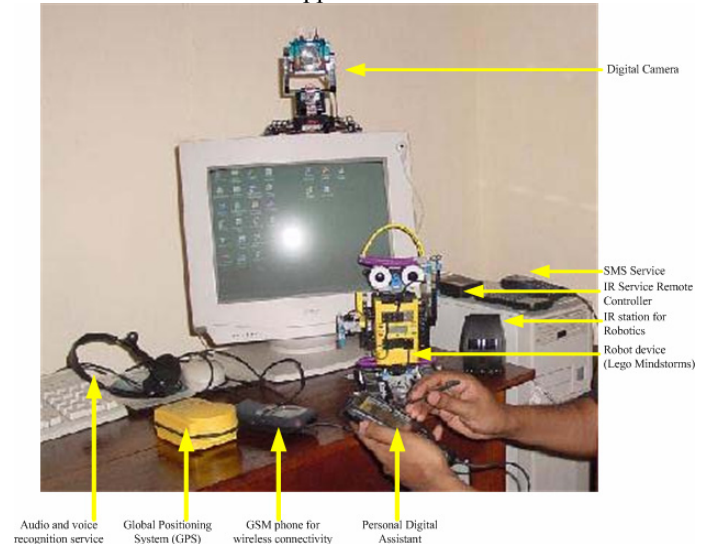


Figure 3: Testing and Simulation Environment for the Intelligent Room Project.

We have tested the current implementation on a limited functionality “Intelligent Room” project to condition the environment state depending on user behaviour. We developed devices to gather environment state information such as motion (user entering or leaving the room and movement to specific parts of the room), temperature, light intensity, noise levels, and speech recognition as primary sensory inputs. The actuations or changing of environment state is carried out by using typical appliances such as lights, air conditioners, fans, televisions, and radio/music players. We initially choreographed episodes of typical human behaviour against a predefined sequence of changing environmental states. Thereafter, collected behavioural patterns exhibited by the AAANTS implementation for the same scenarios. The data gathered showed strong correlation to the expected results.

8. Future Work and Conclusions

We have already performed initial testing using these services to stimulate cooperative behaviour amongst a colony of agents with in a single colony container. The test has shown favourable qualitative results in terms of cooperative behaviour within a single colony with low intensity of real-time sensory signals. We have seen some conflicting behaviour with high intensive environments in terms of activity sequence and correctness. We have also seen duplicate behaviour within a homogeneous community of agents that depict unexpected fuzzy behaviour.

We have devised several methodologies to overcome the deficiencies found in the initial testing phase. We are in the process of enhancing the reinforcement-learning techniques used for stimulating further adaptive behaviour of the agents. In addition, to overcome conflicting behaviour, it is favourable to periodically facilitate evolutionary and reproductive activities to create new flavour of agents and to eliminate individuals that perform poorly over a period of time.

Further, we are in the process of testing the implementation with the smart navigation of a robotic vehicle using computer vision techniques. We are confident that lessons learned from the AAANTS model would contribute to further clarify the understanding of emergent behaviour based on common-sense reasoning.

References

- [1] Martin Beer, Mark d'Inverno, Michael Luck, Nick Jennings, Chris Preist, and Michael Schroeder, 1999, Negotiations in Multi-Agent Systems, A report as a result of a panel discussion at the Workshop of the UK Special Interest Group on Multi-Agent Systems (UKMAS'98)
- [2] Dimitris N. Chorafas, 1998, Agent Technology Handbook, McGraw-Hill, Series on Computer Communications, ISBN-0-07-011923-6.
- [3] Faramarz Farhoodi and Peter Fingar, 1997, Developing Enterprise Systems with Intelligent Agent Technology, Featured in Distributed Object Computing, "DOC" Magazine By Object Management Group (OMG).
- [4] Jacques Ferber, 1999, Multi-Agent Systems, An Introduction to Distributed Artificial Intelligence, Addison-Wesley, ISBN 0-201-36048-9.
- [5] Erich gamma, Richard Helm, Ralph Johnson, John Vlissides, 1995, Design Patterns- Elements of Reusable Object-Oriented Software, ISBN: 81-7808-135-0.
- [6] Michael R. Geneserath, Narinder P. Singh, Mustafa A. Syed, 1994, A Distributed and Anonymous Knowledge Sharing Approach to Software Interoperation, Stanford University.
- [7] Bert Holldobler, Edward O. Wilson, 1990, The Ants, ISBN: 0-674-04075-9
- [8] Timothy Lacey, Scott A. DeLoach, 2000, Automatic Verification of Multi-agent Conversations, Eleventh Annual Midwest Artificial Intelligence and Cognitive Science Conference, University of Arkansas, Fayetteville.
- [9] Timothy Lacey, Scott A. DeLoach, 2000, Verification of Agent Behavioral Models, The 2000 International Conference on Artificial Intelligence (IC-AI'2000) Monte Carlo Resort, Las Vegas, Nevada.
- [10] M. Minsky, 1986, The society of mind, Simon and Schuster, New York, New York. ISBN 0-671-65713-5.
- [11] Hyacinth Nwana & Divine Ndumu, 1996, An Introduction to Agent Technology, Intelligent Systems Research Applied Research and Technology, BT Labs.
- [12] Van Parunak, John Sauter, and Steve Clark, 1997, Toward the Specification and Design of Industrial Synthetic Ecosystems, Fourth International Workshop on Agent Theories, Architectures, and Languages (ATAL'97).
- [13] R.A. Chaminda Ranasinghe, A.P Madurapperuma, 2002, AAANTS – Distributed Mobile Component Architecture for an Ant Colony based Synthetic Ecosystem MATA'02, Fourth International Workshop on Mobile Agents for Telecommunication Applications, Universitat Pompeu Fabra. Barcelona, Spain.
- [14] R.A. Chaminda Ranasinghe, A.P Madurapperuma, 2003, AAANTS-A Distributed Agent Framework for Adaptive and Collective Intelligence, Mobile Agents for Telecommunication Applications (MATA'2003), IEEE/IFIP, Marrakech, Morocco.
- [15] R.A. Chaminda Ranasinghe, A.P Madurapperuma, 2003, Enhanced Frame Based Representation for an Intelligent Environment, (KIMAS'2003) International Conference on Integration of Knowledge Intensive Multi-Agent Systems, IEEE Boston Section, Cambridge MA.
- [16] Narinder P. Singh and Mark A. Gisi, 1995, Coordinating Distributed Objects with Declarative Interfaces, Computer Science Department, Stanford University, Stanford, California, Software Technology Lab, Hewlett-Packard, Laboratories, Palo Alto.
- [17] Peter Stone, 1998, Layered Learning in Multi-Agent Systems, School of Computer Science, Carnegie Mellon University, Pittsburgh.
- [18] Katia Sycara, Matthias Klusch, Seth Widoff, Jianguo Lu, 1999, Dynamic Service Matching Among Open Information Environments, The Robotics Institute, Carnegie Mellon University, U.S.A. Computer Science Department, University of Toronto, Canada.



Engineering Optimisation with Evolutionary Computation

Asanga Ratnaweera , Saman K. Halgamuge, Harry C. Watson*

Mechatronics & Manufacturing research group

*Thermofluids research group

Dept. of Mechanical and Manufacturing Engineering

University of Melbourne, Victoria 3010, Australia.

asangar, sam, hw@mame.mu.oz.au

Abstract- This paper presents a review of various types of evolutionary optimisation methods available for real-world systems optimisation. Initially, we report a brief overview on some evolutionary optimisation methods, their developments and applicability on engineering problems. A comparison of the performance of two different optimisation techniques on well known benchmarks are also presented. A review on real-world applications of evolutionary optimisation methods are also provided.

1 Introduction

In the face of increased global competition and stringent customer expectations, most engineering developments are subjected to challenging performance requirements. Consequently, most real-world developments are driven to reduce costs and time, improve the quality, reliability and efficiency. Therefore, optimisation has become a key aspect in most engineering developments.

Most real-world problems are highly non-linear and often contain combinatorial relationships and uncertainties. Further, they are subjected to many constraints. Therefore, effective mathematical formulation of these problems is challenging and most mathematical models are too complex and subjected to many assumptions. As a result, solutions of classical optimisation methods, such as mathematical programming techniques, can be short of being optimal and may fail to satisfy the feasibility requirements for most real-world problems. Therefore, most of the classical optimisation methods are quite ineffective in solving most engineering optimisation problems.

The recent developments of evolutionary optimisation methods have made the effective optimisation of real-world problems a reality. Most of these optimisation methods operate on the principle of natural evolution or social behaviours of creatures. One of the main advantages of evolutionary optimisation techniques is that they require only a minimum mathematical information about the optimisation problem. Moreover, effectiveness of evolutionary optimisation methods on handling nonlinear problems, defined in continues, discrete and mixed domains, and handling constraints are well documented[7],[21].

Unlike conventional optimisation methods, most evolutionary optimisation techniques are population based: a set of feasible solutions (individuals) are considered in the optimisation process. Further, in contrast to the conven-

tional search optimisation methods, in evolutionary optimisation techniques, individuals exchange their search experiences to move efficiently towards the global optimum solution [2].

Evolutionary optimisation methods start with a random initialisation of individuals in the feasible search space. Then they are manipulated according to a “rule of nature” and, a new population is created at each iteration (generation) until they find the optimum solution or a pre-defined stopping criteria. The manipulation of individuals at each generation is mostly based on the “fitness” of each individual according to an objective function corresponding to the problem being optimised. Therefore, no mathematical information about the behaviour of the problem in the search space is required.

Since the introduction of evolutionary computation concept in 1960’s there have been a few different evolutionary optimisation techniques proposed. Each method has its own method of manipulation of individuals at each generation. Rapid convergence, robustness, flexibility and computational complexity are the key concerns of these different developments. A brief description of some commonly used methods are presented in Section 3.

As a result of the significantly improved features of evolutionary optimisation methods, they have drawn much attention as robust and reliable optimisation methods for real world optimisation problems. Evolutionary optimisation techniques have been applied successfully in combinatorial optimisation, function optimisation, fuzzy logic systems, neural network learning, data mining, bio-computing area, clustering, scheduling problems, engineering designs etc.

In this paper, initially we provide an overview of some commonly used optimisation methods. Some useful recent extensions to the different methods are also provided. The advantages, similarities as well as differences of different methods are also presented. A comparison of the performance of two different methods on two of the well known benchmarks are also presented. A discussion on engineering applications of evolutionary computing is also included.

2 Classification of Optimisation Techniques

There are a lot of optimisation methods available for real-world problem optimisation. However, the performance of most of the methods are highly problem dependent thus the knowledge of the complexity of the optimisation prob-

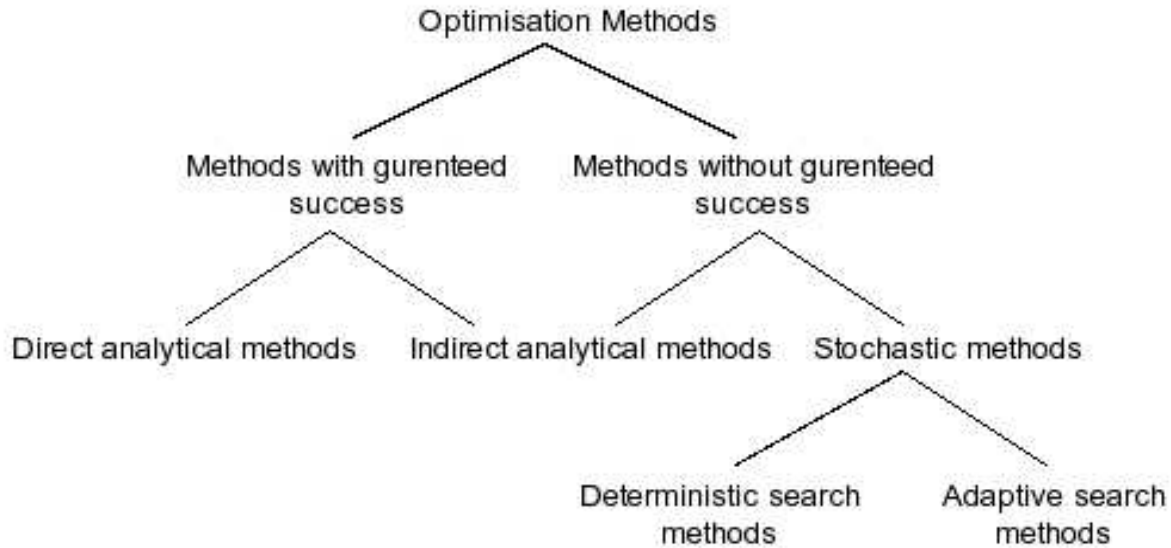


Figure 1: Classification of optimisation methods

lem is extremely important to choose the most efficient method. Generally, optimisation methods are classified according to their functionality. Figure 1 shows a general classification of optimisation methods.

In direct analytical techniques, the search space for the optimisation is discretized and each point is checked for optimisation. Therefore, these methods are highly inefficient for most of the engineering optimisation problems even though they guarantee the optimal solution. On the other hand, analytical methods, such as mathematical programming methods, require a complete mathematical representation of the functionality of the problem. Therefore, these methods are not very effective for most of the complex real-world applications subjected to constraints. However, since some of these methods are capable of finding the global optimum solution, they are quite frequently used for real-world optimisation problems.

In contrast, in stochastic search optimisation methods only a minimum amount of mathematical information is required for the optimisation process. Usually, in these methods, the optimum solution is randomly searched in the feasible search space. Therefore, they do not require mathematical information about the behaviour of the optimisation problem (fitness landscape). However, with these methods there is no guarantee of finding the global optimum solution for most of the complex engineering problems.

In deterministic search methods such as hill climbing methods and greedy algorithms, the input variables are randomly generated and optimisation problem is evaluated (fitness) and tested for the feasibility at each itera-

tion. Most of these methods consider one variable at a time for the function evaluation and the search is driven towards the best solution at each iteration, irrespective of the global information about the fitness landscape. Therefore, with these methods, the possibility of stagnating the search at a local optimum solution is high. Therefore, deterministic search methods are commonly known as local optimisation methods and they are not recommended for multi-variable and multi-modal optimisation problems particularly if the variables are highly correlated and inter-related [41].

Unlike deterministic search algorithms, in most of the adaptive search algorithms, search experience is usually considered to determine the direction of the search and hence the possibility of finding the global optimum solution is significantly high. All evolutionary optimisation methods can be categorised as adaptive search algorithms. Simulated annealing, Tabu search, Evolutionary algorithms, Swarm intelligence can be considered as the most commonly used adaptive search optimisation methods.

Simulated Annealing: Simulated annealing (SA) as its name implies, is based on the physical process of annealing metals; from a high temperature (atoms are randomly distributed) metal cools and freezes into a minimum energy crystalline structure. SA looks for a better solution, in the search space, through a stochastic iterative process with a possibility of accepting a worse configurations. Therefore, SA has the ability to roam in the search space without stagnating at a local optimum solution. Even though,

basic SA can avoid getting stuck at a local optimum, it is found to be not very effective in complex multi-modal landscape [18],[24]. Therefore, as an improvement, multiple initialisation has been proposed. In this strategy, search is carried out from several random initial points in the search space. However, from benchmark simulations, it has been found that SA is significantly slower than most evolutionary methods for most of the complex problems[24].

Tabu Search : Tabu search (TS) is an improved random search technique which considers certain movements in the search space as forbidden or tabu to avoid unnecessary cycling about local optima. As in SA, TS also randomly generates feasible solutions in the neighbourhood of the current solution at each iteration. TS accepts a neighbourhood solution even it is worse than the current best solution [35],[18]. However, to avoid cycling around a previously found best solution, TS creates a list of recently visited points (tabu list) in the search space. If any new solution is listed in the tabu list (recently visited point) the movement to that point is considered as forbidden and movement is avoided. Hence TS avoids getting trapped in a local optimum solution. Further, during the iterative process, TS as SA, always keeps track of the best solution achieved so far. There are several modifications to the basic TS concept have been proposed aiming the efficient performance and rapid convergence to the global optimum solution; aspiration criterion etc.

The major focus of this paper is the evolutionary computation techniques and some of the most commonly used evolutionary optimisation techniques are briefly explained in the following section.

3 Evolutionary Computation Techniques

It has been proven from benchmark simulations that most adaptive search algorithms are capable of solving continues as well as discrete combinatorial optimization problems subjected to linear and nonlinear constraints. Further, these methods have been successfully used to obtain optimal or near optimal solutions for most real-world optimisation problems.

Evolutionary optimisation techniques differ from the other adaptive search optimisation techniques due to two prominent features. Firstly, they all are population based; they consider a population of potential solutions at each iteration. Secondly, the candidate solutions (individuals) communicate each other and exchange information among themselves to find the global optimum solution effectively, within a reasonable amount of iterations (generations). Evolutionary algorithms and Swarm intelligence can be considered as two major types of evolutionary optimisation methods.

Evolutionary Algorithms : All evolutionary algorithms mimic the metaphor of natural biological evolution. Therefore, they work according to Darwin's theory of evolution; survival of the fittest. Hence at each generation, evolutionary algorithms create a new population through selection, competition and recombination. Since the introduction of evolutionary programming in early 60s there has been a few different evolutionary algorithms proposed. Each evolutionary algorithm has its own method of recombination which distinguishes themselves from each other [38].

Swarm Intelligence : Swarm intelligence is a relatively new artificial intelligent concept based on the collective "intelligent" behaviours of "unintelligent" agents. Concept of swarm intelligence is inspired by the collective behaviours of creatures like bees, fish, birds, ants etc. Ant colony algorithms and Particle swarm algorithm are considered as most widely used optimisation techniques based on swarm intelligence.

In the following sub-sections we describe five of the most widely used evolutionary optimisation methods.

3.1 Evolutionary Programming (EP)

Evolutionary programming as an artificial intelligent optimisation concept first introduced by Fogel [39] in 1966. In EP, each individual is represented as a phenotype vector in which each modulus corresponds to the real value of each input variable of the objective function. At each generation, EP applies a mutation (often Gaussian) on each individual to create off-springs. In EP off-springs and parents compete each other for the survival for the next generation. The detailed process of EP is given as follows.

1. Generate an initial population P of size X at random with uniform distribution.
2. Evaluate the fitness f_i of each individual i . (fitness of each individual does not have to be the same as the objective value)
3. Apply mutation to each individual X and produce off-springs X_o as follows,

$$X_o = X + N(0, \sigma^2) \quad (1)$$

Where $N(0, \sigma^2)$ is a Gaussian random number with mean 0 and variance σ^2 .

4. Evaluate the fitness of all the off-springs as in Step 2.
5. Conduct m competitions for each individual in the total population of size $2n$. (A value w_i is assigned to each individual i as follows.

$$w_i = \sum_{t=1}^m w_t \quad w_t = \begin{cases} 1 & \text{if } u \leq \frac{f_r}{f_r + f_i} \\ 0 & \text{otherwise} \end{cases}$$

Where u is a uniformly distributed random number in $[0,1]$ and r is a uniformly distributed random integer in $[0,2n]$.

6. The fittest n number of individuals will survive for the next generation .
7. Go to Step 3 until the stopping criterion is satisfied.

3.1.1 Mutation

As stated above it is common practice to add a Gaussian random number with standard deviation σ to the parents to create offsprings. This process is called mutation. The parameter σ , which is widely known as strategic parameter, provides the necessary self-adaptation of knowledge to proceed the search efficiently. However, improved performance for some problems have been observed with Cauchy random distribution [39].

3.1.2 Selection

In EP usually selection is carried out by competition among individuals according to their fitness as shown in Step 5 above. Individuals with higher fitness has a better chance of survival. However, weak individuals too have a chance for survival which is important to avoid stagnation the search at a local optima.

3.2 Evolutionary Strategies (ESs)

Evolutionary Strategies are also population based optimisation methods operate similar to EP. ESs differ from EP only in the recombination and the selection processes. In ES, deterministic selection methods and recombination operators are used. There are two major types of ESs proposed; $(\mu + \lambda)$ -ES and (μ, λ) -ES, where $(\mu > 1)$ and $(\lambda > 1)$.

In ESs μ number of parents are participated in the recombination process to produce (λ) number of individuals. In $(\mu + \lambda)$ -ES, after the reproduction $(\mu + \lambda)$ individuals are reduced to original size of the population μ by removing the least fit λ individuals. On the other hand, in (μ, λ) -ES, only offsprings are considered for the selection $(\lambda > \mu)$; no individual from the previous generation survive for the next generation. The major steps of ESs are given below.

1. Generate an initial population P of size μ at random with uniform distribution.
2. Select two individuals at random with uniform distribution for recombination.
3. Recombine the parents according to a recombination criteria.

4. Mutate the offsprings as in EP.
5. Repeat Steps 2 to 4 for λ times. (λ offsprings to be created)
6. Select μ number of individuals for the next generation.
 - (a) if $(\mu + \lambda)$ -ES is used
Add the offsprings to the parents and μ individuals are selected.
 - (b) if (μ, λ) -ES is used
From the λ offsprings ($\lambda > \mu$), μ individuals are selected.
7. Repeat Steps 2 to 6 until the stopping criteria is satisfied.

3.2.1 Recombination

In ESs, recombination is used as a strategic parameter as well as a variable control strategy. There are several recombination methods proposed for ESs. Discrete recombination, global discrete recombination, intermediate recombination and in global intermediate recombination concept are the most common recombination strategies widely used with ESs.

3.2.2 Selection

Selection process in ESs is completely deterministic and the best μ individuals from λ or $\mu + \lambda$ are selected according to the ES used.

3.2.3 Mutation

Mutation process in ESs is similar to EP. The Gaussian mutation is commonly used in ESs.

3.3 Genetic Algorithms (GA)

Genetic algorithms were first introduced by Holland in 1960's [14]. As in the other evolutionary optimisation methods, GA are also population based optimisation method work according to the genetic operations of natural evolution [3],[14], [17]. Unlike most other evolutionary optimisation methods, GA operate on genotype level. Therefore, in GA each individual in the population is represented as a chromosome consisting number of genes. GA often find the optimum solution through genetic operations such as crossover and mutation. GA were widely experimented and successfully applied on various real-world optimisation problems in the last few decades. Further, GA often outperform most of the other evolutionary optimisation methods for complex multi-modal problems. The basic operations of genetic algorithms are show below.

1. Generate an initial population $P_{(0)}$ at random.
2. Evaluate the fitness of each individual.

3. Select parents for breeding based on their fitness.
4. Create off-springs through crossover operation.
5. Apply mutation at random.
6. Select individuals for the next generation from off-springs and parents.
7. Repeat Steps 2 to 6 until the stopping criteria is satisfied.

3.3.1 Selection

In evolutionary optimisation methods, selection of individuals is a key aspect in finding the global solution efficiently and effectively. Usually, the natural selection process does not merely select the best candidates for the survival. Instead, potential candidates are chosen statistically such that candidates with higher fitness values have better chance of being selected. Considering these concerns, there are several selection methods introduced; roulette wheel selection, ranked selection and tournament selection are widely used in genetic algorithms [14].

The roulette wheel selection scheme is the simplest and commonly used methods in GA [14]. Under this method, each potential candidate (individual) gets a slot on a roulette wheel whose size is proportional to the relative fitness of the individual. Then the wheel is spun and the individual correspondent to the slot at which the wheel stops is selected. Therefore, this process mimics the natural process of selection by selecting fitter individuals with a chance of selecting weak ones as well.

However, this process has been found to be ineffective particularly when the fitnesses of a few individuals in the population are substantially higher (dominant individuals) than the rest. In such cases selection pressure on the dominant individuals are significantly higher than the rest of the population and as a result the search stagnates around dominant individuals. This phenomena is called genetic drift, which slows down the rate of convergence and leads to premature convergence.

To overcome the problems associated with roulette wheel selection, fitness scaling and rank based selection have been proposed. Fitness scaling is an adjustment of fitness values of each individuals such that differences of the fitness of dominant individuals and weak individuals are small. There are several fitness scaling methods have been proposed [14]. However, the selection of fitness scaling strategy is problem dependent and additional information about the fitness landscape of the problem may be needed to select the appropriate selection method.

In rank based selection individuals are ranked according to their fitness and the rank is used for the selection. However, this strategy is computationally expensive as at each generation individuals have to be sorted according to their fitness. On the other hand, in tournament selection, a number of individuals are selected at random, known as tournament, and the best individual is selected

for the reproduction. Therefore, in this method, the selection pressure is dependent on the tournament size.

3.3.2 Crossover

GA uses the crossover operation to recombine parents to form off-springs. In crossover operation, two of the parents selected in the selection process exchange their “genetic material” (components) to form two off-springs. However, in GA, crossover operation is carried out with a certain probability (crossover probability), often a value close to 1. Therefore, all the parents are not subjected to crossover all the time. There are several different crossover operators proposed. Moreover, selection of the type of the crossover operator as well as the crossover probability are problem dependent. A few well known crossover operators are explained below [14].

Single point crossover : Initially a point (crossover point) is selected at random and each parent is split into two segments at this point. Then the off-springs are created by swapping the segments between parents as shown below.

Before crossover

parent 1 :	00000000 11111111111111
parent 2 :	11111111 00000000000000

After crossover

offspring 1 :	00000000000000000000
offspring 2 :	11111111111111111111

Multi-point crossover : In multi-point crossover strategy, several crossover points are selected at random on each chromosome (parent) and off-springs are created by swapping corresponding segments between two parents. Therefore, multi-point crossover provides more mixing of genetic materials than single point crossover but this may be disruptive for some problems due to excessive mixing.

Before crossover

parent 1 :	0000 000000 00000 111111
parent 2 :	1111 111111 11111 0000000

After crossover

offspring 1 :	00001111110000000000
offspring 2 :	11110000001111111111

Uniform crossover : The uniform crossover is usually done by randomly shuffling the genes of the parents. Therefore, in this strategy, each gene of parents is considered as a crossover point. Hence, each gene of an offspring is copied from the corresponding gene of one of the parents with equal probability. This is usually done by defining a random mask chromosome (binary) with equal number of genes as parents. Then, each gene of the parents are swapped according to the corresponding gene value of the mask. Uniform crossover strategy provides rapid mixing of genetic materials and adds the necessary diversity to find

the optimum solution. However, this can be disruptive for some problems and may not converge to the optimum solution within reasonable amount of generations.

Before crossover

parent 1 :	0000000000000001111111
mask :	10101000101010101011
parent 2 :	11111111111111110000000

After crossover

offspring 1 :	1010100010010100101100
offspring 2 :	01010111010101010111

3.3.3 Mutation

Mutation process normally adds new genetic materials to chromosomes and hence provide additional diversity. This is usually done by changing one of the genes of a chromosome of an offspring with a new value; in binary coded GA mutation is done by swapping a bit. Off-springs are selected for mutation according to a certain probability (mutation probability) which is often a small value compared to the crossover probability. Mutation has been identified as an important process to avoid premature convergence of GA to a local optimum solution [14].

3.4 Particle Swarm Optimisation (PSO)

Particle swarm concept was first introduced as an effective function optimisation strategy by Kennedy and Eberhart [22] in 1995. Since then there have been a lot of experimental investigations done on PSO and significant improvements have been made on the original PSO concept [32], [21],[6],[1],[10].

PSO is inspired by animal social behaviours such as flock of birds and school of fish. In the natural process, creatures like fish and birds act as a swarm to find the food sources efficiently and protect themselves from the predators. Analogous to this animal social behavior, PSO finds the optimum solution by considering the search experiences of each particle in the swarm as well as exchanging the search experience between particles.

In PSO, initially a swarm of particles (candidate solutions) are generated at random and manipulated them in the search space by changing the magnitude and the direction of the velocity of each particle. Therefore, the search towards the global optimum is governed by the velocity calculation for each particle. Generally, the velocity of each particle is calculated at each generation according to the following equation.

$$\tilde{\mathbf{v}}_i = \mathbf{w} \times \tilde{\mathbf{v}}_i + \mathbf{C}_1 \times \mathbf{rand}_1() \times (\tilde{\mathbf{p}}_i - \tilde{\mathbf{x}}_i) + \mathbf{C}_2 \times \mathbf{rand}_2() \times (\tilde{\mathbf{p}}_g - \tilde{\mathbf{x}}_i) \quad (2)$$

Where $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{v}}_i$ are the position vector and the velocity vector of a particle i at time t . Further, $\tilde{\mathbf{p}}_i$ and $\tilde{\mathbf{p}}_g$ are the position vectors associated with the best solution achieved by the particle i and the current global best solution in the

swarm respectively. The constants C_1 and C_2 are called acceleration coefficients and \mathbf{rand}_1 and \mathbf{rand}_2 are two uniformly distributed random numbers between [0,1]. w is called inertia weight factor. At each time step (generation) next position of the particle i is updated according to the following equation.

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{v}}_i + \tilde{\mathbf{x}}_i \quad (3)$$

The first part of the equation 2 represents the contribution of the previous velocity to the new velocity. This component provides the necessary momentum for particles to roam around the search space without stagnating in a local optimum. The second part, known as cognitive component, provides the local information for the search by incorporating search experience of each particle to the velocity calculation. Finally, the third component, also called social component, represents the global thinking of each particle and adds the global search information of the swarm to each particle. The basic operation of PSO is as follows.

1. Generate an initial population $P(0)$ at random.
2. Evaluate the fitness of all the individuals according to a user defined objective function.
3. Estimate the new velocity for each particle according equation (2).
4. Move the particles for the next function evaluation according equation (3).
5. Repeat Steps 2 to 4 until the stopping criteria is satisfied.

There are many studies reported on the effect of each parameter of equation (2) on the performance of PSO through benchmark simulations [31],[6],[36],[29],[9], [30]. Eberhart and Shi [9] found that it is desirable to reduce the inertia weight factor from a higher value to a lower value during the search. Further, they have found that a random inertial weight factor is useful for tracking dynamic systems and constraint handling [11]. In addition, in our previous studies, we observed an improved performance for some benchmarks, with time varying acceleration coefficients alone with time varying inertia weight factor [28]. Further, we studied about the possible fuzzy adaptation of acceleration coefficients as well [27].

It has been reported that for most commonly used test function, PSO works better than most other evolutionary optimisation methods. Considering the superior performance for most problems and relative simplicity of implementation, PSO has drawn much attention as a robust global optimisation method for real-world system optimisation in the last few years.

3.5 Ant Colony Algorithm (ACA)

As its name implies, ACA is based on the colonization behaviour of ants. ACA mimics the ants behaviour of establishing the shortest route paths from their colony to food sources and back [8]. It has been observed that each ant lays a chemical substance which is called “pheromone”, on the ground when they are walking towards a food source. The intensity and the quantity of pheromone is found to be dependent on the distance to an food source and the desirability of the food. Usually, ants in the colony follow the pheromone trails laid by the others. Investigations have shown that ants always prefer to follow a pheromone trail which extrudes the strongest scent. Hence, the colony slowly settles down to the shortest path, which is the strongest pheromone trail.

This collective behaviour of ants has been interpreted as an optimisation method by Dorigo [8] in early 1990’s. A probabilistic model has been developed to mimic the colonization behaviour of ants. Ant colony algorithm is found to be an effective optimisation method for distributed systems optimisation such as travelling salesmen problem.

4 Parameter Control of Evolutionary Computation

The performance of all the evolutionary algorithms are found to be significantly sensitive to their control parameters such as population size, mutation and crossover probabilities, selection methods and so on. On the other hand, most of the parameters are found to be significantly problem dependent. Therefore, in order to make the maximum use of these algorithms appropriate parameter control is a challenging task.

There are a few parameter automation strategies have been developed and self-adaptive parameter controlling have drawn much attention in the last few years, [12],[19], [5], [20],[31]. Use of fuzzy control techniques for self-adaptation of parameters are considered as an efficient strategy for most complex problems [12],[19],[16],[17]. Further, compared to the most evolutionary algorithms, particle swarm technique has less number of control parameters (inertia weight factor and acceleration coefficients) [33].

5 Performance of Evolutionary Optimisation Methods

Performance of different evolutionary optimisation methods have been studied through benchmark simulations. There are a lot of comparative studies of different algorithms have been reported [23],[1], [29]. However, in most of these studies only the performance of a basic version of one algorithm is compared with advanced developments of the other method. Therefore, in this study, performance of two different GA strategies are compared with two different PSO strategies.

5.1 Benchmarks Simulation

Benchmark simulation is widely used to evaluate and compare the performance of optimisation methods. Further, most advancements on optimisation methods are often done through benchmark simulation. In this study, we compare the performance two different GA strategies and two different PSO strategies. Two of the well known benchmarks with different complexities, are considered in this study. One function (Rosenbroke function) is a uni-modal function whereas the other one (Rastrigrin function) is a multimodal function. Both functions have inter-correlated variables. Mathematical representation of the functions are as follows.

Rosenbroke function

$$f_1 = \sum_{i=1}^n [100(x_{i+1} - x_i)^2 + (x_i - 1)^2] \quad (4)$$

Rastiring function

$$f_2 = \sum_{i=1}^n [x_1^2 - 10\cos(2\pi x_i) + 10] \quad (5)$$

5.2 Simulation strategies

In this comparative study, the number of function evaluations is kept the same for all the methods. All the simulations were carried out with population size of 40. Both function were considered in 15 dimensions. All the simulations were forced to stop after 8×10^4 function evaluations. Individuals are initialised symmetrically in the search space at random. Ranges of initialisation and the limits of the search for each function are given in Table 2.

A description of different GA and PSO strategies used are as follows.

Table 1: different GA and PSO strategies used

Method	Description
GA1	simple GA with roulette wheel selection and 100% replacement.
GA2	steady state GA with linear scaling and tournament selection (50% of the population is replaced at each generation).
PSO1	Fixed acceleration coefficients at 1.494 and random inertia weight ($w=0.5 + \text{rand}()/2$), where $\text{rand}()$ is uniformly distributed random number in $[0,1]$.
PSO2	time varying inertial weight and acceleration coefficients. Inertia weight is changed from 0.9 to 0.4 whereas the acceleration coefficients C_1 and C_2 are changed from 2.5 to 0.5 and 0.5 to 2.5 respectively.

Table 2: Operating and initialisation ranges

Function	initialisation range	operating range
f_1	$(-100, 100)^n$	$(-100, 100)^n$
f_2	$(-5.12, 5.15)^n$	$(-5.12, 5.15)^n$

5.3 Results

The results show the importance of the control parameters of both GA and PSO. For both functions, both PSO strategies showed improved mean optimum solution when compared to the two GA strategies used. Further, GA2 showed improved mean optimum solution over GA1 for the function f_1 but its performance, in terms of mean optimum solution, on function f_2 shown to be comparatively poor. However, GA1 has found a better minimum solution for 20 trial when compared to GA2. On the other hand, PSO1 performed significantly well on function f_1 but PSO2 outperformed PSO1 for function f_2 in terms of the mean optimum solution. Therefore, the results clearly show the significance of selecting an appropriate optimisation method as well as control parameters for problem optimisation.

Table 3: Maximum, mean and optimum values for 20 trials

Function	GA1	GA2	PSO1	PSO2
f_1	147.17	124.27	145.21	136.25
	52.27	25.83	15.77	28.09
	2.27	13.18	0.01	0.06
f_2	26.31	35.71	17.35	14.93
	14.19	25.5	11.42	8.65
	7.83	18.75	4.808	4.97

6 Engineering Applications of Evolutionary Optimisation Methods

Evolutionary optimisation techniques have been applied successfully to many real-world optimisation problems as well as to problems that can be converted to optimisation problems. Some of the most potential application areas are production planning and scheduling, manufacturing systems, robotic applications, decision making, telecommunication networks, electrical power systems, neural networks, image processing, classification and biological systems. A few prominent applications of evolutionary optimisation techniques are briefly described in this section.

Most real-time complex distributed computing systems are generally NP-hard and often conventional optimisation methods have difficulties in finding feasible solutions. However, evolutionary optimisation methods have shown promise on various scheduling problems ranging from Job-shop scheduling problems to scheduling tasks in multi-processor systems [15],[16],[40], [4], [10]. Further, these optimisation methods have been successfully applied on time table problems and routing table optimisation [34].

In addition, evolutionary optimisation methods have successfully been applied for load scheduling in electrical power generation systems and hence significant improvement of economic efficiency of power generation has been achieved[13].

Evolutionary computation techniques have been extensively applied on neural network based applications [25], [14]. Further, they have been used to evolve weights of neural networks as well as their structure. Moreover, modified evolutionary computation techniques are also reported in order to make these algorithms feasible for real time learning of neural networks [4]. Further, possible use of evolutionary optimisation methods in on-line adaptive learnable revolvable hardware was also reported [26].

In pattern recognition applications, it is often desirable to obtain the minimum feature set for maximum classification accuracy. Evolutionary computation techniques have been successfully applied for feature selection and identification in pattern recognition and classification with significantly better classification accuracy when compared with conventional techniques [3].

Robotics is another promising area of application of evolutionary optimisation. Evolutionary computing techniques have been used for controlling and motion planning of robot manipulators and mobile robots. These techniques along with the other soft computing techniques such as neural networks and fuzzy control, have been used to identify dynamic model for multi-link robot manipulators and for behaviour based control and path planning of robots [37].

Most of the complex engineering design problems consist of a lot of input parameters and subjected to a lot of constraints. Therefore, selection of the best set of parameters is often challenging. Most of the complex structural and mechanical design problems have been successfully solved using evolutionary optimisation methods [7]. Further, in our investigation on applying evolutionary optimisation methods to select the design and operating parameters of internal combustion engines, significant performance improvement when compared to the conventional optimisation methods have been observed.

7 Conclusions

The major focus of this paper was to provide a comprehensive review on the evolutionary computation techniques and there applications to real world optimisation problems. Initially we explained the operating methodologies of some of the commonly used evolutionary optimisation methods. Then we provided some simulation results comparing the performance of genetic algorithms and particle swarm algorithm. Effect of control parameters of these two algorithms on the performance was also presented. Finally we provided a brief overview on different real-world applications of the evolutionary computational techniques [18].

Acknowledgments

The first author, on leave from the University of Peradeniya, Sri Lanka, is funded by the Asian Development Bank project on Science and Technology Personnel Development, under the Ministry of Science and Technology, Sri Lanka, and the Melbourne International Research Scholarship of University of Melbourne.

Bibliography

- [1] P. J. Angeline. Evolutionary optimization verses particle swarm optimization: Philosophy and the performance difference. *Proceedings of the 7th International Conference on Evolutionary Programming, Lecture notes in computer science, 1447, Evolutionary Programming VII*, 2:600 – 610, March 1998.
- [2] M. badami, M. R. Marzano, and P. Nuccio. Influence of late intake valve opening on the s.i. engine performance in idle condition. *SAE international congress & Expositions Detroit Michigan*, (960586), Feb. 1996.
- [3] N. Chaiyaratana and A. Zalzala. Recent developments in evolutionary and genetic algorithms: theory and applications. *Second International Conference On Genetic Algorithms In Engineering Systems: Innovations And Applications*, pages 270 – 277, 1997.
- [4] S. O. Chang and J. K. Lee. New approach to real-time adaptive learning control of neural networks based on an evolutionary algorithm. *Proceedings of IEEE International Symposium on Industrial Electronics*, pages 1871 – 1876, 2001.
- [5] M. Clerc. The swarm and the queen: Towards a deterministic and adaptive particle swarm optimization. *Proceedings of the IEEE International Congress on Evolutionary computation*, 3:1951 – 1951, March 1999.
- [6] M. Clerc and J. Kennedy. The particle swarm: explosion, stability and convergence in a multi-dimensional complex space. *IEEE Trans. on Evolutionary Computation*, 6:58 – 73, Feb. 2002.
- [7] R. Dinger. Engineering design optimization with genetic algorithms. *Conference Proceedings North-con/98*, pages 114 – 119, 1998.
- [8] M. Dorigo. The ant colony optimisation meta-heuristic : Algorithms, applications and advances, technical report iridla-2002-32. World Wide Web, <http://www.agent.aitia.ai/download.php?ctag=download&docID=360>, 1998.
- [9] R. C. Eberhart and Y. Shi. Comparing inertia weights and constriction factors in particle swarm optimization. *Proceedings of the IEEE International Congress on Evolutionary computation*, 1:84 – 88, 2000.
- [10] R. C. Eberhart and Y. Shi. Particle swarm optimization: Developments, applications and resources. *Proceedings of the IEEE International Conference on Evolutionary Computation*, 1:81 – 86, 2001.
- [11] R. C. Eberhart and Y. Shi. Tracking and optimizing dynamic systems with particle swarms. *Proceedings of IEEE Congress on Evolutionary Computation, Piscataway, NJ, Seoul, Korea*, 2:94–97, 2001.
- [12] A. E. Eiben, R. Hinterding, and Z. Michalewicz. Parameter control in evolutionary algorithms. *IEEE Trans. on Evolutionary Computation*, 3(2):124–141, 1999.
- [13] A. Gaul, E. Handschin, and W. Hoffmann. Evolutionary strategies applied for an optimal management of electrical loads. *Proceedings of International Conference on Intelligent Systems Applications to Power Systems*, pages 368 – 372, 1996.
- [14] D. E. Goldberg. *Genetic Algorithms in search, Optimization, and Machine Learning*. Addison & Wesley, Reading MA, 1989.
- [15] G. Greenwood, C. Lang, and S. Hurley. Scheduling tasks in real-time systems using evolutionary strategies. *Proceedings of the IEEE Third Workshop on Parallel and Distributed Real-Time Systems*, pages 195–196, 1995.
- [16] G. Greenwood and A. G. K. McSweeney. Scheduling tasks in multiprocessor systems using evolutionary strategies. *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, pages 345–349, 1994.
- [17] J. J. Grefenstette. "optimisation of control parameters for genetic algorithms". *IEEE Trans. on Systems, man and Cybernetics*, 1986.
- [18] J. Hao and J. Pannier. Simulated annealing and tabu search for constraint solving, artificial intelligence and mathematics iv. (1998). World Wide Web, <http://www.citeseer.nj.nec.com/hao98simulated.html>, 1998.
- [19] F. Herrera and M. Lozano. Fuzzy genetic algorithms: Issues and models. World Wide Web, <http://www.citeseer.nj.nec.com/16799.html>.
- [20] R. Hinterding, Z. Michalewicz, and A. E. Eiben. Adaptation in evolutionary computation: A survey. *Proceedings of the 4th IEEE International Conference on Evolutionary Computation*, April 1997.
- [21] X. Hu, R. C. Eberhart, and Y. Shi. Engineering optimisation with particle swarm. *Proceedings of the IEEE Swarm Intelligence Symposium 2003 (SIS 2003), Indianapolis, Indiana, USA*, pages 53–57, 2003.

- [22] J. Kennedy. The particle swarm:social adaptation of knowledge. *Proceedings of the IEEE International Conference on Evolutionary Computation*, pages 303–308, 1997.
- [23] J. Kennedy and W. M. Spears. Matching algorithms to problems: An experimental test of the particle swarm and some genetic algorithms on the multimodal problem generator. *Proceedings of IEEE World Congress on Computational Intelligence*, 2:78 ? 83, May 1998.
- [24] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.
- [25] S. Kremer. Genetic algorithms for protein tertiary structure prediction. *IEE Colloquium on Applications of Genetic Algorithms*, pages 6/1 –6/5, 1994.
- [26] D. W. Lee, C.-B. B. K. B. Sim, H.-S. S. K.-J. Lee, and B.-T. Zhang. Behavior evolution of autonomous mobile robot using genetic programming based on evolvable hardware. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 3835 –3840, 2000.
- [27] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson. Particle swarm optimisation with self adaptive acceleration coefficients. *Proceedings of the International conference on Fuzzy Systems and Knowledge Discovery 2002, Singapore*, 2002.
- [28] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson. Particle swarm optimisation with time varying acceleration coefficients. *Proceedings of the International conference on Soft Computing and Intelligent Systems 2002, Tsukuba, Japan*, 2002.
- [29] Y. Shi and R. C. Eberhart. Comparison between genetic algorithms and particle swarm optimization. *Proceedings of the 7th International Conference on Evolutionary Programming, Lecture notes in computer science, 1447, Evolutionary Programming VII*, 1:611 – 616, March 1998.
- [30] Y. Shi and R. C. Eberhart. A modified particle swarm optimizer. *Proceedings of the IEEE International Conference on Evolutionary computation*, 1:69 – 73, 1998.
- [31] Y. Shi and R. C. Eberhart. Parameter selection in particle swarm optimization. *Proceedings of the 7th International Conference on Evolutionary Programming, Lecture notes in computer science, 1447, Evolutionary Programming VII*, 1:591 – 600, March 1998.
- [32] Y. Shi and R. C. Eberhart. Empirical study of particle swarm optimization. *Proceedings of the IEEE International Congress on Evolutionary computation*, 3:101 – 106, 1999.
- [33] Y. Shi and R. C. Eberhart. Fuzzy adaptive particle swarm optimization. *Proceedings of the IEEE International Congress on Evolutionary computation*, 1:101 – 106, March 2001.
- [34] M. Sinclair. The application of a genetic algorithm to trunk network routing table optimisation. *Proceedings of 10th Teletraffic Symposium on Performance Engineering in Telecommunications Network*, pages 2/1 –2/6, 1993.
- [35] O. Steinmann, A. Strohmaier, and T. Sttzle. Tabu search vs. random walk.
- [36] P. N. Suganthan. Particle swarm optimizer with neighborhood operator. *Proceedings of the IEEE International Congress on Evolutionary computation*, 3:1958 – 1962, 1999.
- [37] K. Watanabe and K. Izumi. A survey of robotic control systems constructed by using evolutionary computations. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 758 –763, 1999.
- [38] X. Yao. Global optimisation by evolutionary algorithms. *Proceedings of the Second AIZU International Symposium on Parallel Algorithms/Architecture Synthesis, pages 282-291, Aizu-Wakamatsu, Japan, 17.-21. March 1997. IEEE Computer Society Press, Los Alamitos, CA. yCCA33980/97 ga97aXYao.*, 1997.
- [39] X. Yao, Y. Liu, and G. Lin. Evolutionary programming made faster. *IEEE Trans. on Evolutionary Computation*, 3:82–102, July 1999.
- [40] W. Ying and L. Bin. Job-shop scheduling using genetic algorithm. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 1994 –1999, 1996.
- [41] D. Yuret and M. de la Maza. "dynamic hill climbing: Overcoming the limitations of optimization techniques". *The Second Turkish Symposium on Artificial Intelligence and Neural Networks*, 1993.

Workshops/Tutorials

How to build a successful e-government

Outline

Electronic Government, or “E-Government,” is one of the main interested topics in field of ICT and Sri Lanka. E-government is designed to make better use of information technology (IT) investments to eliminate wasteful government spending, reduce government’s paperwork burden on citizens and businesses, and improve government response time to citizens – from months down to minutes. A key goal of successful e-government is for citizens to be able to access government services and information within three “clicks,” when using the Internet. This workshop consists of two comprehensive presentations that cover the various important topics in the area of e-government.

Resource Persons

Prof. Takefuji, Keio University, Japan

Mr. Lalith Weeratunga, ICT Agency, Sri Lanka

Prof. V.K. Samaranayake, University of Colombo School of Computing, Sri Lanka and National Center for Digital Governments, Harvard University, USA

Duration Half Day

Intended Audience

Personnel such as Directors, CEO’s, Managers and Information Technology specialists responsible for planning, implementing or managing e-government application from both public and private sector.

EU-Asian IT&C stakeholder meeting

Outline

Propose of this workshop is to give an idea of the scope of the projects implemented under the EU-Asia IT&C program. EU-Asia IT&C is designed to foster economic growth and understanding between Europe and Asia through better awareness, access to, and use of Information and Communication Technologies (ICT). Attending this workshop will provide information on what EU-Asia IT&T is, example of projects already underway and how participant can apply for EU support to work in partnership with European institution to develop program in Asia.

Resource Persons

Mr. Laurent Vanopstal, Asian Secretariat, Asia IT&C Program, Bangkok, Thailand

Prof. V.K. Samaranayake, Member, Advisory panel, Asia T&C and Director, University of Colombo School of Computing, Sri Lanka

Duration Half Day

Intended Audience

Directors, CEO's and Managers with interest in ICT project from European government grants and other non-profit organizations

Sponsored by EU-Asia IT&C program

E-Commerce Security – Problems, Solutions and Trends**Outline**

This tutorial will provide a comprehensive overview of problems, solutions and trends in the area of electronic commerce security. The tutorial highlights the risks posed by insecure e-commerce systems and identifies strategies, which help to mitigate these risks. The tutorial examines the relevant Internet security protocols and communication standards such as SET and SSL, cryptographic schemes, key management strategies, authentication mechanisms, browser vulnerabilities, Web server security, privacy concerns, digital certificates and PKI and provides practical insights.

Resource Persons

Jeffy Mwakalinga, Researcher, Royal Institute of Technology, Sweden.

Chih-Chun Chang, Researcher, George Washington University, USA.

Kasun De Zoysa, Lecturer, University of Colombo School of Computing, Sri Lanka.

Rasika Dayarathne, Lecturer, University of Colombo School of Computing, Sri Lanka.

Duration Full Day

Intended Audience

IT professionals responsible for planning, implementing or managing a secure e-commerce infrastructure including secure payment processing, secure web servers, and supporting web technologies. Principles outlined in this tutorial are equally applicable to both educational and commercial organizations in the public and private sector.

Developing Secure Java Applications**Outline**

This tutorial will provide a comprehensive overview of developing secure Java applications. Experience Java security application developers will lead you through important areas of Java security including Signed Applets, Java Cryptography, Secure RMI and Java Smart Cards. This tutorial offers Java developers, a unique opportunity to fine-tune their knowledge and skills in the area of Java security.

Resource Persons

Jeffy Mwakalinga, Researcher, Royal Institute of Technology, Sweden.

Chih-Chun Chang, Researcher, George Washington University, USA.

Kasun De Zoysa, Lecturer, University of Colombo School of Computing, Sri Lanka.

Rasika Dayarathne, Lecturer, University of Colombo School of Computing, Sri Lanka.

Duration Full Day

Intended Audience

Java application developers and other software developers intending to move into secure electronic commerce, e-government and banking applications.

Planning and implementation of switched-fixed-wire line and wireless networks

Outline

With the proliferation of low cost-high speed Ethernet technology, it has now become possible to implement efficient campus networks, both small scale and large scale. In this tutorial we will look at the switched Ethernet technology and the IP technology, the steps to taken from planning stage to the implementation stage of such networks and wireless local area technologies, and last mile options. This tutorial will cover the topics such as Switch network design, Layer 2 and Layer 3 switching, Megabit and gigabit Ethernet, VLANs and IEEE 802.1q tagging, Provision of QoS for integrated traffic, Routing and the role of routers in a switched infrastructure, IP management with NAT and PAT, Wireless LANs and 802.11x specification and integration, Bluetooth technology, Tender specification for fiber and UTP cabling, Implementation, Testing and Commissioning of networks and use of OTDR, Service level agreement and Bandwidth management.

Resource Persons

Dr. Buddy Liyannage, B.Sc., Ph.D., MIEE, CCNA, Independent Consultant, UK

Dr. D.N. Ranasinghe, B.Sc. (Eng), Ph.D., MIEE, Senior Lecturer, University of Colombo School of Computing, Sri Lanka.

Duration Full Day

Intended Audience

Those who are about to embark on a new network infrastructure project, looking to enhance and improve exiting network installations. The spread of topics would appeal to wide audience of differing experience and knowledge.

Upgrade your business with e-learning

Outline

Good e-learning promotes communication and makes collaborative problem solving possible at any time and any place in an organization. This workshop will discuss how this can be achieved by using real life examples. It will answer questions like; how can I use e-learning in a innovative way to upgrade my organization? Why do we need computers to start communicating

with each other's? Why is learning design so important? What about tools and standards? The tutorial offers attendees a unique opportunity to fine-tune their skills and acquire new ones by presenting a combination of real life examples, expert knowledge, discussions and hands-on training.

Resource Persons

Dr. Johan Torbionnsson, Stockholm University, Sweden

Mrs. Margareta Hellstrom, The Swedish Netuniversity, Sweden

Duration Half Day

Intended Audience

Small & Medium sized Enterprises (SMEs), persons engaged in Human Resource Development (HRD), Sales & Marketing and / or Communication & Information activities in the ICT community (or in any other organization).

Sponsored by Sida and e-learning centre of University of Colombo School of Computing

Upgrade your teaching with e-learning

Outline

This workshop will discuss how e-learning and the use of ICT can improve the way we teach and learn by using real life examples. It will answer questions like; how can I use e-learning in an innovative way to upgrade my teaching? What happens when we put the learner and learning in the centre? Why is the teacher's role so important for student's learning? The tutorial offers attendees a unique opportunity to fine-tune their skills and acquire new ones by presenting a combination of real life examples, expert knowledge, discussions and hands-on training.

Resource Persons

Dr. Johan Torbionnsson, Stockholm University, Sweden

Mrs. Margareta Hellstrom, The Swedish Netuniversity, Sweden

Duration Half Day

Intended Audience

Teachers in Higher education, in tertiary education, schools teachers, persons engaged in Human Resource Development (HRD) in the government or community (or in any other organization).

Sponsored by Sida and e-learning centre of University of Colombo School of Computing

Tutorial on Bio Informatics

Outline

Bioinformatics is a fast growing field within the biological sciences that was developed because of the need to handle large amounts of genetic and biochemical data. This data, originating from individual research efforts, is linked by its common origin: *the cell of living organism*. To understand the links between pieces of information from research areas such molecular biology, structural biochemistry, cell biology, physiology and pathology, the bioinformatics uses computational power to catalogue, organize, and structure these pieces into biologically meaningful entities.

As Information and Communication Technology dominated the latter part of the 20th century, it is said that the 21st century belongs to the biotechnology. With high speed computing power at one's disposal, it is now possible to work on some of the major problems in Biotechnology linked to the Human genome project and other genome sequences currently in the public domain. These areas are fashionably referred to as Genomics, Proteomics and Structural Biology. This tutorial covers major areas of the bio-informatics with hands-on experience in on-line databases in Genetics.

Resource Persons

Dr. Siv Andersson, Uppsala University, Sweden

Dr. Ranil Dassanayake, Department of Biochemistry & Molecular Biology, University of Colombo, Sri Lanka

Dr. Jagath Weerasena, Department of Biochemistry & Molecular Biology, University of Colombo, Sri Lanka

Dr. Nalin de Silva, Department of Chemistry, University of Colombo, Sri Lanka

Dr. Preethi Randeniya, Department of Zoology, University of Colombo, Sri Lanka

Dr. Shiroma Handunetti, Malaria Research Unit, Sri Lanka

Dr. Ruwan Weerasingha, University of Colombo School of Computing, Sri Lanka

Mr Harasha Wijewardena, University of Colombo School of Computing, Sri Lanka.

Duration Full Day

Intended Audience

Biomedical sciences and Computer Scientists

Sponsored by Sida

Implementation of UNICODE compatible Sinhala language support

Outline

Internet has become so essential for the development of a country; all nations developed and developing are trying their best to encourage their citizenry to use Internet. One of the major impediments of the use of the Internet has always been that the Internet is fuelled by Roman Script based languages.

One of the standardize solution for the computerization in languages other than roman script based is Unicode. The intended workshop will guide the participants from new development in Unicode Sinhala to Unicode Keyboard drivers with hand on experience in tools such as PFAEDIT, VOLT etc. Topics such as Salient features of Unicode, How local languages fits into Unicode, Sinhala and Unicode, Unicode Wijeesekera Keyboard, Unicode fonts using VOLT and PFAEDIT, Unicode fonts with Microsoft and Opensource software will be coved in this tutorial.

Resource Persons

Dr. Ruwan Weerasinghe, Senior Lecturer, University of Colombo School of Computing Sri Lanka

Mr. Harsha Wijayawardhana, Consultant, University of Colombo School of Computing Sri Lanka

Duration Full Day

Intended Audience

Aimed at people who are involved in developing software application with local language support.

The Current and future IT innovative equipment**Outline**

In this tutorial attendees will hear about the newest advancements in IT related equipment. Presenter will introduce and demonstrate various IT equipment and the innovations that add value to it. At the end of this tutorial participants may understand where the IT equipment industry is moving and how will future concerns be addressed.

Resource Persons

Prof Takefuji, Keio University, Japan

Intended Audience

Aimed at people who are interested to get innovative experience on IT equipments

Sponsored by Sida

Computational Intelligence in the real-world**Outline**

The major objective of this tutorial is to assemble a collection of presentations that reflect the latest advances in the area of artificial intelligence. These presentations cover both the recent advancement of artificial intelligence techniques and their applications in the field of engineering and science. Topics such as Evolutionary computation techniques, Fuzzy control systems, Neural

networks and their applications, Data mining, Hybrid soft computing techniques, Computer vision and image processing, and Machine learning will be covered in this tutorial.

Resource Persons

Prof Saman Halgamuge, University of Melbourne, Australia
Mr. Ravindra Koggalage, University of Melbourne, Australia
Mr. Asanga Ratnaweera, University of Melbourne, Australia
Dr. Ruwan Werasinghe, Senior Lecturer, University of Colombo School of Computing, Sri Lanka.

Duration Full Day

Intended Audience

Those who are interest in designing, implementing, and deploying artificial intelligence based systems

Tutorial on Web services**Outline**

Web services use a sophisticated infrastructure to provide a simple mechanism for client applications to invoke methods and obtain results from server applications regardless of differences in source languages and host platforms. This tutorial will cover various topics in the area of web services and simulate a common use case: taking an existing service, and exposing it as a Web service for point-to-point synchronous integration. It provides an opportunity to understand the process of creating, deploying, and testing a Web service – processes shared by most full-scale applications.

Resource Persons

Dr. Sanjeva Weerawarna, IBM and Lanka Open Source Foundation

Intended Audience

Those who are interest in designing, implementing, and deploying web-based business systems

Broadband Access**Outline**

One of the most exciting developments that the world of communications has witnessed in recent times is broadband access technologies. This tutorial will offer a technical overview of recent developments in this area through presentations and demonstration. Topics such as Digital Subscriber Loop (DSL), Data over Cellular Systems, 3rd Generation Cellular Systems, Fixed Broadband Wireless Access, and Wireless LANs will be covered during the presentation. Therefore, participant of this tutorial will obtain a good understanding of the modern broadband communication technologies and their applications, and how each technology can contribute to the development of the telecommunication infrastructure in Sri Lanka.

Resource Persons

Dr. Deleeka Dias, University of Moratuwa, Sri Lanka
Dr. Gihan Dias, University of Moratuwa, Sri Lanka
Dr. Nandana Rajatheva, University of Moratuwa, Sri Lanka
Mr. Kithsiri Samarasinghe, University of Moratuwa, Sri Lanka

Duration Full Day

Intended Audience

Telecommunication service providers, Users of modern communication services who are interested in learning how broadband access work, and how they can be used to improve the productivity of their organizations.

ICT for Blind People**Outline**

This workshop teaches and demonstrates Digital Accessible Information System (DAISY) for blind people. DAISY is often used to refer to a standard for producing accessible and navigational multimedia documents for those people who are blind or visually impaired. Using DAISY standard, content creators, library serving people and book publisher can produce a Digital Talking Book (DTB), synchronize an electronic text file with an audio, generate electronic Braille file from electronic text and produce a structured digital “text-only” documents for blind and visually impaired personnel. This tutorial offers a wide variety of presentations based on above for audiences who have different levels of knowledge about the DAISY standard and its applications.

Resource Person

Mr. Kawamura Hiroshi, Keio University, Japan

Duration Three Days

Intended Audience

People who are blind or visually impaired and digital content creators, library serving people, book publisher and instructors who work for blind or visually impaired people.
[Sponsorship is available for limited number of participants]

Sponsored by JICA and ADMTC, University of Colombo School of Computing

Business Process reengineering**Outline**

Business Re-engineering or Business Process Re-engineering (BPR) is a widely used or misused term, often means different things to different people. There are almost as many methodologies

for BPR as there are active consultants in the field. At the workshop it is expected to discuss various theoretical frameworks available in the current literature which are formulated to cover various aspects of BPR process, especially the application of IT and then to illustrate the need for integrative and holistic approach for success. In addition, it is expected to discuss many case examples and to cover the application of methodologies, tools and techniques of BPR to practical situation.

Duration Full Days

Resource Person

Dr.Bandu Ranasinghe, B.Sc,MBA,Ph.D,FCS,MACS,FBCS, IDM Group of Company, Sri Lanka

Intended Audience

IT manages, Application developers



Infotel Lanka Society

Council Members

Chairman	Prof. V.K. Samaranayake
President/SLCVA	Mr. Hyder Alaudeen
President/SLASI	Mr. Sriyan De Silva
Chairman/BCS	Mr. Jayantha De Silva
President/LISPA	Mr. Rohith Udulagama
President/ISACA	Mr. N. Supramaniam
Chairman/Finance Committee	Mr. K. Ramathas

IITC 2003 Main Organizing Committee

Chairman

Prof. V.K. Samaranayake
Mr. Chandana Weerasinghe
Dr. Sanjiva Weerawarana
Dr. Bandu Ranasinghe
Mr. Sriyan De Silva
Dr. Prasad Wimalaratne

Finance

Mr. K. Ramathas
Mr. David Dominic
Mrs. C. Jayalath

Papers & Workshops

Dr. Ruvan Weerasinghe
Dr. Ajith Madurapperuma
Dr. Carmel Wijegunawardena
Dr. Ajantha Athukorala
Dr. Kasun De Zoysa
Dr. Rukshan Athauda

Publicity

Mr. Rohith Udulagama
Mr. Rohan Sirisena
Mr. Niranjana De Silva

Registration

Mrs. Girty Gamage



Strategic Partner

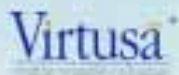


Sri Lanka Telecom

Principal Sponsor



Co-sponsors:



B-S-S



LankaCom
Services

Subsidiary of Sri Lanka Telecom International Limited



ENTERPRISE SOLUTIONS



UCSC